FUSIONCLIP: MULTIMODAL CONTRASTIVE PRETRAINING FOR FUSION PLASMA DIAGNOSTICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Visible-light videos and multivariate diagnostic signals provide complementary observations of tokamak discharges. However, existing fusion pretraining methods, primarily learn representations by reconstructing or predicting observations, leaving the correspondence between these modalities underexplored. Representation learning is further challenged by the fact that plasma-related regions occupy only part of each video frame and diagnostic variables have substantially different sampling rates. We propose FusionCLIP, a multimodal contrastive pretraining framework. Spatiotemporal Stability Selection retains informative visual tokens, while a Temporal Query Encoder integrates asynchronous diagnostic signals using their observation timestamps. Contrastive learning then aligns the two modality representations. We construct EAST Multimodal Diagnostics (EAST-MMD), comprising 1307 EAST discharges, and evaluate the model on three independently fine-tuned tasks: disruption prediction, edge-localized mode (ELM) classification, and H_{98} regression. FusionCLIP achieves an aggregate score of 89.07%, exceeding the best general-purpose pretraining configuration and an in-domain contrastive pretraining configuration without Spatiotemporal Stability Selection by 6.27 and 2.12 percentage points, respectively. Ablations and attention visualizations further support the contributions of its key components.

1 INTRODUCTION

Nuclear fusion has the potential to provide low-carbon energy, while stable tokamak operation requires continuous monitoring of high-temperature plasmas. During a tokamak discharge, different diagnostic systems observe the same plasma process from complementary perspectives. As shown in Figure 1(a), visible-light videos capture plasma radiation structures and their spatial evolution, whereas multivariate diagnostic signals record the temporal evolution of different physical quantities. Learning transferable plasma representations from these observations is an important problem for intelligent fusion diagnostics. Pretraining offers a way to learn transferable representations from large-scale data. In fusion research, FusionMAE (Yang et al., 2026) and TokaMind (Boschi et al., 2026) have demonstrated the value of pretrained representations for downstream tasks through the reconstruction or prediction of observed data. Videos and diagnostic signals from the same time window form natural pairs, motivating cross-modal contrastive pretraining (Radford et al., 2021) to learn representations transferable across fusion diagnostic tasks. However, the distinct characteristics of the two modalities, pose challenges to learning such representations.

As shown in Figure 1(a), plasma-related regions occupy only part of each frame amid extensive background and repeated observations. Studies of visual representation learning suggest that how local information contributes to a representation matters. Registers (Darcet et al., 2024) showed that background tokens may be repurposed for internal computation, while LaSt-ViT (Shi et al., 2026) demonstrated that selective aggregation affects visual representations. Existing methods for vision-based fusion diagnostics (Chen et al., 2026a; Wang et al., 2026) primarily extract spatiotemporal features for specific diagnostic tasks. The role of local-region selection in learning transferable video representations remains less explored. Therefore, we explicitly select local information during video encoding, allowing the details of selected regions to be retained more fully. For diagnostic time series, a central challenge is the heterogeneity of sampling frequencies. The signals in Figure 1(a) range from 25 Hz to 250 kHz. Existing fusion models often resample signals onto a fixed temporal

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.

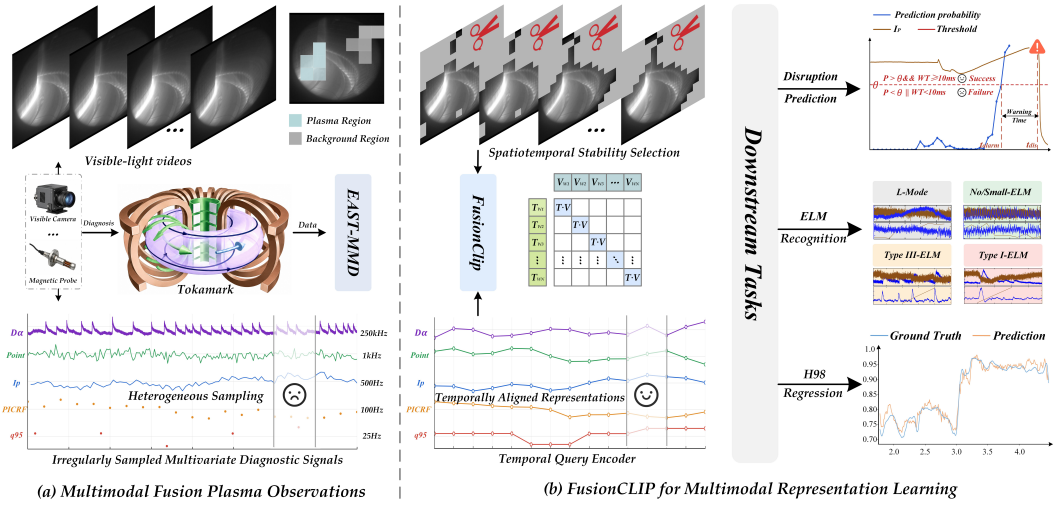

Figure 1: Overview of EAST-MMD and FusionCLIP. (a) Paired plasma videos with redundant backgrounds and diagnostic signals with heterogeneous sampling rates. (b) Spatiotemporal Stability Selection and Temporal Query Encoder encode the two modalities for contrastive pretraining, followed by independent fine-tuning for disruption prediction, ELM classification, and H_{98} regression.

grid (Yang et al., 2023). A fixed grid simplifies input construction but may compress high-frequency variations and over-interpolate low-frequency signals, obscuring their original temporal structure.

To address these challenges, we propose FusionCLIP, a multimodal contrastive pretraining framework. As shown in Figure 1(b), the visual branch uses Spatiotemporal Stability Selection(SSS) to select tokens, preserving local details while reducing interference from background and repeated observations. Its temporal branch employs a Temporal Query Encoder (TQE), which uses temporal queries and time-biased attention to transform heterogeneously sampled signals into aligned representations. The two branches are jointly pretrained with a bidirectional contrastive objective.

We construct the EAST Multimodal Diagnostics (EAST-MMD) dataset and evaluate FusionCLIP on three independent tasks: disruption prediction, edge-localized mode (ELM) classification, and H98 regression. Compared with approaches using existing pretrained visual backbones, FusionCLIP achieves better results on multiple diagnostic metrics.

Our main contributions are as follows:

- We introduce FusionCLIP, a multimodal contrastive pretraining framework for paired plasma videos and diagnostic signals.
- We introduce SSS and Temporal Query Encoder to construct modality-specific representations that address the uneven distribution of local information in videos and the heterogeneous sampling of diagnostic signals, respectively.
- We construct EAST-MMD and evaluate the proposed pretraining framework and its key components on three independent diagnostic tasks.

2 Related Work

2.1 Pretraining Paradigms

Pretraining has become an important approach to learning transferable representations from large-scale unlabeled or weakly supervised data. Based on their primary training objectives, multimodal pretraining methods encompass masked modeling and reconstruction (Geng et al., 2022; Wang et al., 2023; Chen et al., 2024), autoregressive modeling (Wang et al., 2024), cross-modal contrastive learning (Radford et al., 2021; Jia et al., 2021), and combinations of multiple objectives (Li et al., 2021;

Bao et al., 2022b). These methods learn transferable representations through complementary training signals, including recovering observed content, predicting sequential dependencies, and aligning paired samples.

Recent work has also explored pretrained representations for fusion plasmas. FusionMAE learns a plasma-state representation from multichannel HL-3 diagnostics through compression-based and missing-signal reconstruction (Yang et al., 2026). TokaMind models heterogeneous MAST observations, including time series, profiles, and videos, and learns transferable representations through multi-signal reconstruction and prediction objectives (Boschi et al., 2026). Bouchiat et al. (2025) presented an initial exploration of masked autoencoding for integrating multi-diagnostic and control data from DIII-D. Together, these efforts demonstrate the potential of learning reusable representations from large-scale fusion observations.

Representative fusion pretraining efforts have primarily formulated their objectives around reconstructing or predicting observations. The correspondence between videos and diagnostic signals acquired within the same time window provides another training signal for representation learning, as suggested by contrastive pretraining in other multimodal domains (Radford et al., 2021; Jia et al., 2021). Contrastive pretraining based on paired plasma videos and multivariate diagnostic signals, and its transfer to fusion diagnostic tasks, remains underexplored. We therefore propose FusionCLIP to study multimodal contrastive pretraining by aligning representations of paired plasma videos and diagnostic signals.

2.2 Visual and Temporal Representation Learning

Visual representation learning has increasingly considered the quality of local features and their aggregation beyond image-level performance. DINO (Caron et al., 2021), Registers (Darcet et al., 2024), and LaSt-ViT (Shi et al., 2026) highlight the role of local information organization from the perspectives of spatial semantics, background tokens, and selective aggregation, respectively. Temporal representation learning must likewise account for the sampling structure of observations. For regularly sampled sequences, TCN (Bai et al., 2018) and PatchTST (Nie et al., 2023) model temporal dependencies on a fixed time grid. For missing or irregular observations, GRU-D (Che et al., 2018) and mTAN (Shukla & Marlin, 2021) incorporate observation intervals or timestamps into the encoding process. Together, these studies underscore the importance of organizing local visual information and modeling observation times when learning representations.

In fusion diagnostics, visible-light videos capture plasma morphology and evolution, yet salient plasma-related regions occupy only part of the field of view. Existing studies typically use neural networks to extract spatial and temporal features from videos (Chen et al., 2026a; Wang et al., 2026). These methods primarily learn features for specific diagnostic tasks. The role of local region selection in learning transferable video representations remains underexplored. Fusion diagnostic signals often differ substantially in sampling frequency. Existing approaches can be broadly grouped into two categories. The first organizes signals on a fixed time grid (Yang et al., 2023; Croonen et al., 2023). The second retains differences in sampling rates across diagnostics. The cross-tokamak disruption prediction study (Zheng et al., 2023) uses diagnostic-specific sampling rates and time windows, while FusionMMT (Chen et al., 2026b) models heterogeneous signals with multi-scale, time-aware, and variable-aware representations.

Motivated by these studies, FusionCLIP employs spatiotemporal structure-guided token selection in its visual branch to preserve details from plasma-related regions. Its temporal branch uses time-aware queries to encode asynchronous, multi-rate diagnostic signals. The resulting representations support contrastive pretraining between paired videos and diagnostic signals.

3 EAST Multimodal Diagnostics Dataset

3.1 Dataset Curation

As shown in Fig. 1(a), we construct EAST Multimodal Diagnostics (EAST-MMD), comprising visible-light videos and multivariate diagnostic signals from 1307 EAST discharges. 667 discharges are used for multimodal pretraining and 640 for downstream evaluation. For pretraining, we use 50ms windows with a 2ms sliding stride, yielding 2725021 paired video-signal windows. The

visible-light camera operates at 1kHz, and 12 frames are uniformly sampled from each window. The diagnostic input contains 38 variables, each retaining its original observation times and sampling rate without resampling onto a common grid. Downstream evaluation uses the task annotations and data splits of EAST-VTD640 (Chen et al., 2026b). This dataset contains 481 non-disruptive and 159 disruptive discharges, yielding 280094 annotated windows. The training, validation, and test sets are split by discharge and contain 447, 96, and 97 discharges, respectively.

3.2 Downstream Tasks

As illustrated in Fig. 1(b), we evaluate the pretrained representations on disruption prediction, ELM classification, and H_{98} regression.

Disruption Prediction. The model predicts disruption risk from the visible-light video and diagnostic signals in each window. At test time, windows from the same discharge are processed chronologically, and the first prediction to exceed a threshold selected on the validation set determines the alarm time. For a disruptive discharge, an alarm is counted as successful only if it precedes the disruption by at least 10ms. A threshold crossing on a non-disruptive discharge is counted as a false alarm.

ELM Classification. The model identifies the plasma edge state associated with each window. The four classes are L-mode, No/Small ELM, Type-III ELM, and Type-I ELM.

H_{98} **Regression.** The model predicts a continuous H_{98} value for each window to characterize the plasma energy confinement level.

Further details on the dataset, diagnostic signals, and task labels are provided in Appendix A.1.

4 FusionCLIP

4.1 Overview

For a given 50-ms discharge window, let $\mathbf{V}_i$ denote the visible-light video and let the multivariate diagnostic time series be $O_i = \{(o_{i,m,j}, t_{i,m,j})\}_{m,j}$, where $o_{i,m,j}$ is the j-th observation of diagnostic variable m and $t_{i,m,j}$ is its actual timestamp. As shown in Figure 2, FusionCLIP extracts representations using a video encoder E_v and a time-series encoder E_o. The video branch employs Spatiotemporal Stability Selection to organize local visual information, while the temporal branch uses a Temporal Query Encoder to read observations at their respective timestamps. Their outputs are denoted by

$$\mathbf{h}^{(v)}_{\mathrm{CLS},i} = E_v(\mathbf{V}_i), \qquad \mathbf{h}^{(o)}_{\mathrm{CLS},i} = E_o(\mathbf{O}_i). \tag{1}$$

The two representations are used for multimodal contrastive pretraining and subsequently transferred to downstream diagnostic tasks.

4.2 Structure-Guided Video Encoding

The video encoder first converts consecutive frames into spatiotemporal patch tokens and uses Transformer interactions to form local features. Spatiotemporal Stability Selection combines feature stability, spatial consistency, and temporal persistence. The selected tokens retain their individual representations, while the unselected tokens are aggregated into summary tokens. The CLS output is then used as the global video representation $\mathbf{h}^{(v)}_{\mathrm{CLS}}$. Let $\mathbf{h}^{(v)}_{\tau,n} \in \mathbb{R}^D$ denote the visual feature at spatial position n in temporal patch τ. The normalized ranks of the three components are denoted by $R^F_{\tau,n}$, $R^S_{\tau,n}$, and $R^T_{\tau,n}$, respectively. As illustrated in Figure 2, Rank fusion produces the final selection score:

$$P_{\tau,n} = 0.20R^F_{\tau,n} + 0.35R^S_{\tau,n} + 0.45R^T_{\tau,n}. \tag{2}$$

The encoder selects high-scoring positions within each temporal patch.

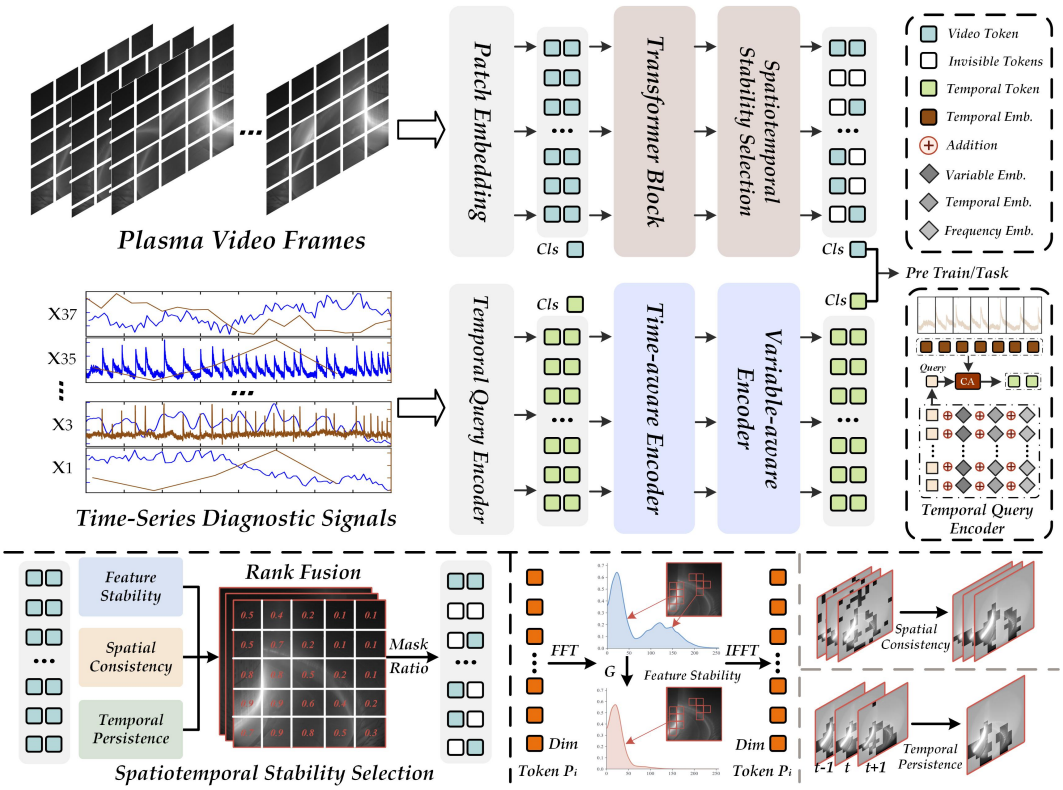

Figure 2: Overview of FusionCLIP. Spatiotemporal Stability Selection selects video tokens using feature stability, spatial consistency, and temporal persistence, while Temporal Query Encoder maps irregularly sampled diagnostic signals onto a shared temporal grid.

Feature Stability. Self-attention in the video encoder allows each patch feature to incorporate information from other regions. We use a low-pass perturbation to examine how strongly each patch depends on high-frequency feature components. A response supported mainly by smoother components should remain relatively similar after filtering, whereas a response that relies more on scattered high-frequency components should change more substantially. The difference between the original and filtered responses therefore provides a stability cue for selecting local information. We expect tokens containing substantial plasma-related information to exhibit more consistent responses across feature dimensions, preserving their main information under the low-pass perturbation. In contrast, background regions may combine unrelated details and be more sensitive to the suppression of high-frequency components. We compare each patch feature before and after filtering to guide local information selection.

Specifically, we apply a Fourier transform along the feature dimension of $\mathbf{h}_{\tau,n}^{(v)}$, suppress high-frequency components with G, and transform the result back:

$$\widetilde{\mathbf{h}}_{\tau,n}^{(v)} = \mathrm{irFFT}\Big(G \odot \mathrm{rFFT}(\mathbf{h}_{\tau,n}^{(v)})\Big), \qquad G_k = \exp\left[-\frac{1}{2}\left(\frac{k}{\sqrt{D}}\right)^2\right]. \tag{3}$$

As the frequency index k increases, G_k decreases and progressively attenuates high-frequency components along the feature dimension. For feature dimension d, we compute the response relative to the filtering-induced change:

$$u_{\tau,n,d} = \frac{h_{\tau,n,d}^{(v)}}{\max\Big(\Big|\widetilde{h}_{\tau,n,d}^{(v)} - h_{\tau,n,d}^{(v)}\Big|, \epsilon\Big)}. \tag{4}$$

Within the same feature dimension, positions with smaller changes after filtering tend to receive higher scores, indicating that they may preserve more plasma-related information. At each temporal

index τ, we compare the stability responses of the N spatial patches separately for each feature dimension d. Each dimension votes for the patch with the highest response, and $F_{\tau,n}$ denotes the fraction of votes received by patch n:

$$n^*_{\tau,d} = \arg\max_{1\le m\le N} u_{\tau,m,d}, \qquad F_{\tau,n} = \frac{1}{D}\sum_{d=1}^{D} \mathbb{I}[n^*_{\tau,d} = n]. \tag{5}$$

Ranking $F_{\tau,n}$ across the spatial patches at temporal index τ yields the feature-stability rank $R^F_{\tau,n}$.

Spatial Consistency. As shown in Figure 2, feature voting can highlight plasma-related patches, but high-scoring positions may remain spatially scattered. Since visible plasma structures often extend across neighboring regions, we apply Gaussian spatial smoothing to the vote map: $S = \mathcal{G}(\mathbf{F})$, where $\mathcal{G}$ denotes Gaussian neighborhood smoothing. Ranking $\mathbf{S}$ yields the spatial-consistency rank R^S.

Temporal Persistence. Plasma structures evolve throughout a discharge, so an isolated response at a single temporal index is insufficient to characterize temporal persistence. We aggregate responses at the same spatial position across neighboring temporal patches using weights in a $1:2:1$ ratio:

$$C_{\tau,n} = S_{\tau-1,n} + 2S_{\tau,n} + S_{\tau+1,n}. \tag{6}$$

Ranking $\mathbf{C}$ yields the temporal-consistency rank R^T. This component favors positions that receive support across neighboring time steps.

Finally, SSS selects visual tokens according to the fused score $P_{\tau,n}$. Unselected tokens within each temporal patch are averaged into one summary token. In subsequent encoding, the CLS token reads from the selected tokens and summary tokens, retaining local detail in selected regions while incorporating context from the remaining regions.

4.3 Query-Based Temporal Encoding

Fusion diagnostic variables are sampled at substantially different frequencies. Resampling them onto a fixed grid may suppress high-frequency variations or introduce extensive interpolation for low-frequency signals. To preserve observation-time information, we use a Temporal Query Encoder (TQE) to construct the time-series representation.

TQE encodes low-frequency variables observation by observation and high-frequency variables patch by patch. Patch-based encoding first reduces the number of events produced by densely sampled signals and the cost of subsequent processing. Let $\mathbf{p}_{i,m,k}$ denote the k-th observation patch of a high-frequency variable, and let $\tilde{t}$ denote the time associated with an event, derived from its observation timestamps. The two types of event features are

$$\begin{aligned} \mathbf{e}^{(L)}_{i,m,j} &= f_L(o_{i,m,j}) + \phi(\tilde{t}_{i,m,j}) + \mathbf{v}_m + \mathbf{c}_L, \\ \mathbf{e}^{(H)}_{i,m,k} &= f_H(\mathbf{p}_{i,m,k}) + \phi(\tilde{t}_{i,m,k}) + \mathbf{v}_m + \mathbf{c}_H. \end{aligned} \tag{7}$$

Here, f_L and f_H map individual observations and observation patches, respectively. The function ϕ is a time embedding, $\mathbf{v}_m$ is a variable embedding, and $\mathbf{c}_L$ and $\mathbf{c}_H$ are event-type embeddings. Below, $\mathbf{e}_{i,m,r}$ and $\tilde{t}_{i,m,r}$ denote either type of encoded event and its associated time.

Within each 50-ms window, TQE uses 25 learnable time queries shared across variables. Adding the query time s_ℓ and a variable embedding gives

$$\mathbf{q}_{i,m,\ell} = \mathbf{q}_\ell + \phi(s_\ell) + \mathbf{v}_m, \qquad \ell = 1, \dots, 25. \tag{8}$$

Each query attends to the events of its corresponding variable. Although both queries and events contain time embeddings, attention may still assign high weights to temporally distant observations with similar features. We therefore add a time-distance bias to encourage attention to nearby signal changes. For attention head h, the logit is

$$a^{(h)}_{i,m,\ell r} = \frac{(\mathbf{W}^{(h)}_Q \mathbf{q}_{i,m,\ell})^\top (\mathbf{W}^{(h)}_K \mathbf{e}_{i,m,r})}{\sqrt{d_h}} - \lambda_h \left| s_\ell - \tilde{t}_{i,m,r} \right|, \qquad \lambda_h = \mathrm{softplus}(\gamma_h), \tag{9}$$

Table 1: Comparison of different visual backbones on three downstream tasks.

Method	Pre-training	Disruption					ELM			H98		Joint ↑
		TPR ↑	FPR ↓	F1 ↑	AUC ↑	AWT (ms) ↑	ACC ↑	F1 ↑	AUC ↑	MAE ↓	R^2 ↑	
Scratch	–	91.67%	16.44%	75.86%	94.29%	1293.95	89.43%	73.46%	96.46%	0.034669	90.10%	79.81%
BEiT	ImageNet-1K	79.17%	15.07%	70.37%	89.61%	383.79	88.95%	71.62%	96.17%	**0.025314**	93.18%	78.39%
DINO	ImageNet-1K	75.00%	**2.74%**	81.82%	93.44%	310.06	89.86%	73.55%	96.45%	0.026186	93.04%	82.80%
MoCo v3	ImageNet-1K	75.00%	12.33%	70.59%	89.55%	402.89	89.22%	73.41%	96.67%	0.028204	92.40%	78.80%
iBOT	ImageNet-1K	83.33%	9.59%	78.43%	92.92%	385.15	89.18%	73.29%	96.45%	0.030527	91.88%	81.20%
EVA	ImageNet-1K	83.33%	9.59%	78.43%	90.95%	389.15	89.51%	74.08%	96.51%	0.027868	92.18%	81.56%
PixMIM	ImageNet-1K	87.50%	13.70%	76.36%	93.01%	464.33	89.79%	74.07%	96.70%	0.031755	91.30%	80.58%
MAE	ImageNet-1K	79.17%	6.85%	79.17%	89.47%	296.16	90.11%	75.38%	96.70%	0.026656	92.67%	82.41%
CLIP	WIT-400M	87.50%	16.44%	73.68%	94.15%	**1340.67**	88.53%	72.46%	96.34%	0.030182	91.96%	79.37%
MetaCLIP	MetaCLIP	79.17%	9.59%	76.00%	90.84%	381.68	89.68%	73.90%	96.68%	0.029458	92.54%	80.81%
EVA-CLIP	EVA-CLIP	83.33%	6.85%	81.63%	93.09%	316.65	89.32%	74.12%	96.44%	0.028896	92.07%	82.61%
CLIP-Base	EAST-MMD	91.67%	5.48%	88.00%	95.49%	149.41	**91.56%**	79.18%	**97.82%**	0.025787	**93.68%**	86.95%
Ours	EAST-MMD	**95.83%**	**2.74%**	**93.88%**	**99.54%**	222.17	90.92%	**79.82%**	97.61%	0.025362	93.50%	**89.07%**

where d_h is the attention-head dimension and λ_h controls the strength of the time-distance bias. Attention weights are normalized over valid events, allowing variables with different sampling rates to form representations at shared query times. A temporal backbone then models dependencies within each variable, followed by cross-variable attention. Finally, the resulting features are aggregated into the global representation $\mathbf{h}^{(o)}_{\mathrm{CLS},i}$.

4.4 Multimodal Contrastive Pretraining

The video and diagnostic time series from the same window, $(\mathbf{V}_i, \mathbf{O}_i)$, form a positive cross-modal pair. We project the global representations of the two encoders into a shared space. For $a \in \{v, o\}$, let $\mathbf{z}^{(a)}_i = \mathrm{norm}(\mathbf{W}_a \mathbf{h}^{(a)}_{\mathrm{CLS},i})$ be the normalized projection. The cross-modal similarity is $d_{ij} = (\mathbf{z}^{(v)}_i)^\top \mathbf{z}^{(o)}_j / \theta$, where the learnable temperature θ is initialized to 0.07.

Let $\mathcal{I}$ denote the set of valid paired samples and $\mathcal{A}_i$ the candidate set containing the positive pair. Invalid samples and other windows that overlap with window i within the same discharge are excluded as negatives. The video-to-time-series and time-series-to-video losses are

$$\mathcal{L}_{v\to o} = -\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \log \frac{e^{d_{ii}}}{\sum_{j\in\mathcal{A}_i} e^{d_{ij}}}, \qquad \mathcal{L}_{o\to v} = -\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \log \frac{e^{d_{ii}}}{\sum_{j\in\mathcal{A}_i} e^{d_{ji}}}. \tag{10}$$

The pretraining objective averages the two directions:

$$\mathcal{L}_{\mathrm{CLIP}} = \frac{\mathcal{L}_{v\to o} + \mathcal{L}_{o\to v}}{2}. \tag{11}$$

This objective trains the encoders to capture the correspondence between paired videos and diagnostic time series. After pretraining, the encoders are transferred to downstream diagnostic tasks and fine-tuned independently.

5 Experiments

5.1 Evaluation Metrics and Experimental Setup

Evaluation Metrics. The three downstream tasks have different output forms, and we adopt task-specific evaluation metrics. For disruption prediction, we use discharge-level evaluation and report TPR, FPR, F1, AUC, and average warning time (AWT). For ELM classification, we report ACC, Macro-F1, and Macro-AUC. For H_{98} regression, we use MAE and R^2. The definitions of AWT and the overall Joint metric are given as:

$$AWT = \frac{1}{N_s}\sum_{i=1}^{N_s}(t^i_{dis} - t^i_{alarm}), \qquad Joint = \frac{F1_{dis} + F1_{elm} + R^2_{H98}}{3}. \tag{12}$$

Here, N_s is the number of disruptive discharges with a successful warning. Higher TPR, F1, AUC, AWT, and R^2, as well as lower FPR and MAE, indicate better performance.

Table 2: Ablation study of multimodal pretraining, SSS, and TQE.

Pre-training	SSS	TQE	Disruption					ELM			H98		Joint↑
			TPR↑	FPR↓	F1↑	AUC↑	AWT (ms)↑	ACC↑	F1↑	AUC↑	MAE↓	R^2 ↑	
✗	✗	✗	**95.83%**	5.48%	90.20%	98.00%	**326.83**	88.75%	69.88%	96.22%	0.032112	91.60%	83.89%
✗	✗	✓	83.33%	**2.74%**	86.96%	98.06%	73.85	**90.95%**	76.25%	97.81%	0.026874	93.13%	85.44%
✗	✓	✓	**95.83%**	4.11%	92.00%	97.43%	96.39	90.57%	77.21%	**97.86%**	0.030168	92.23%	87.15%
✓	✓	✓	**95.83%**	**2.74%**	**93.88%**	**99.54%**	222.17	90.92%	**79.82%**	97.61%	**0.025362**	**93.50%**	**89.07%**

Table 3: Effects of visual token selection and mask ratio on diagnostic performance.

Mask Ratio	Disruption					ELM			H98		Joint ↑
	TPR ↑	FPR ↓	F1 ↑	AUC ↑	AWT (ms) ↑	ACC ↑	F1 ↑	AUC ↑	MAE ↓	R^2 ↑	
Time Series	87.50%	9.59%	80.77%	94.09%	396.00	90.76%	75.63%	97.42%	0.026838	92.66%	83.02%
Full Token	91.67%	5.48%	88.00%	95.49%	149.41	91.18%	79.18%	97.82%	0.025787	93.68%	86.95%
Random 50%	91.67%	5.48%	88.00%	96.00%	60.05	90.36%	77.67%	97.46%	0.034071	91.09%	85.59%
SSS 25%	79.17%	2.74%	84.44%	96.52%	156.11	90.99%	79.08%	97.76%	0.023024	94.11%	85.88%
SSS 50%	95.83%	2.74%	93.88%	99.54%	222.17	90.92%	79.82%	97.61%	0.025362	93.50%	**89.07%**
SSS 75%	87.50%	5.48%	85.71%	96.38%	149.38	90.14%	77.61%	97.29%	0.026311	93.22%	85.51%

Experimental Setup. We pretrain FusionCLIP on EAST-MMD for 10 epochs using AdamW with a batch size of 256, a learning rate of 3×10^{-5}, and a weight decay of 0.05. Each sample contains 12 video frames and diagnostic signals from 38 variables, with a visual token mask ratio of 50%. Contrastive learning uses a 256-dimensional embedding and a temperature of 0.07. We fine-tune the pretrained model separately for each downstream task for 10 epochs, using a batch size of 128 and a learning rate of 5×10^{-5}, while keeping the weight decay and mask ratio unchanged. Pretraining uses two NVIDIA A800 GPUs, and fine-tuning uses one.

5.2 Comparison with Pretrained Models

Table 1 compares training from scratch, general-purpose pretrained visual backbones, and EAST-MMD-pretrained models. The general-purpose group includes image self-supervised models BEiT (Bao et al., 2022a), DINO (Caron et al., 2021), MoCo v3 (Chen et al., 2021), iBOT (Zhou et al., 2022), EVA (Fang et al., 2023), PixMIM (Liu et al., 2024), and MAE (He et al., 2022), alongside image–text models CLIP (Radford et al., 2021), MetaCLIP (Xu et al., 2024), and EVA-CLIP (Sun et al., 2023). These baselines use TCN (Bai et al., 2018) for diagnostic signals. CLIP-Base and FusionCLIP instead use TQE and are pretrained on paired EAST-MMD observations, with SSS used only by FusionCLIP.

DINO leads the general-purpose group with an aggregate score of 82.80%, compared with 86.95% for CLIP-Base and 89.07% for FusionCLIP. FusionCLIP performs particularly well on disruption prediction (93.88% F1, 99.54% AUC, and 2.74% FPR). It also achieves the highest ELM macro-F1 (79.82%) and near-best H98 regression results, although CLIP-Base has slightly higher ELM accuracy and AUC. Overall, FusionCLIP provides the strongest aggregate performance across the three tasks.

5.3 Ablations and Analysis

Pretraining and Component Analysis. Table 2 shows that sequentially adding TQE, SSS, and contrastive pretraining raises the aggregate score from 83.89% to 85.44%, 87.15%, and 89.07%. TQE improves ELM F1 and H98 MAE, while SSS raises disruption F1 from 86.96% to 92.00%. Pretraining further improves the key metrics across all three tasks, with disruption F1 reaching 93.88%, ELM F1 reaching 79.82%, and H98 MAE falling to 0.025362.

Visual Token Analysis. As shown in Table 3, all multimodal settings achieve higher aggregate scores than the time-series-only baseline of 83.02%, indicating that video provides complementary information. SSS at a 50% mask ratio achieves an aggregate score of 89.07%, exceeding full tokens at 86.95%, random selection at the same mask ratio at 85.59%, and the other SSS mask ratios. Although the 25% SSS setting performs best on H98 regression, the 50% setting offers a better overall balance across the three tasks. We therefore use a 50% mask ratio by default.

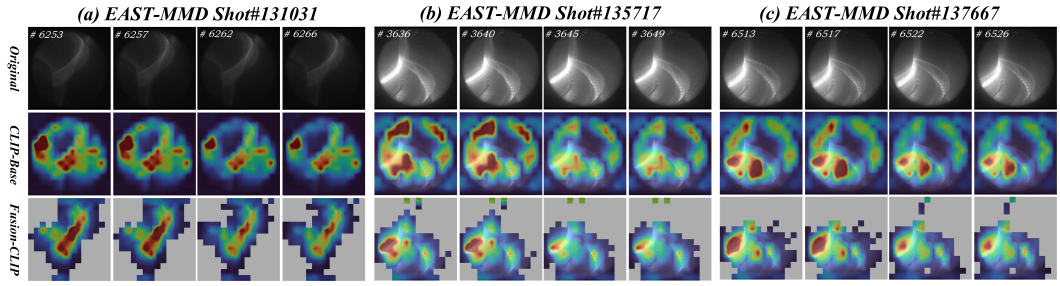

Figure 3: CLS-to-patch attention distributions of pretrained CLIP-Base and FusionCLIP models on EAST-MMD. The first row shows the original video frames. The second and third rows show the pretrained attention-weight maps of CLIP-Base and FusionCLIP, respectively.

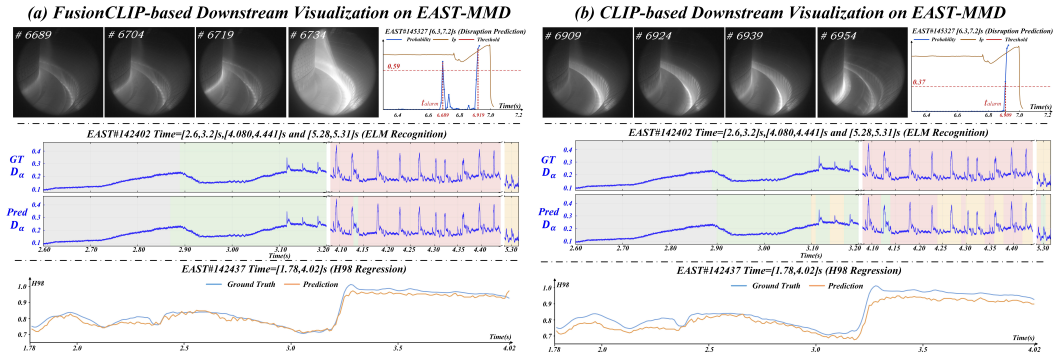

Figure 4: Downstream prediction examples for disruption prediction, ELM classification, and H98 regression on EAST-MMD. (a) Predictions from FusionCLIP. (b) Predictions from CLIP-Base on the corresponding examples.

5.4 Visualization Analysis

Pretraining Attention Visualization. Figure 3 compares the CLS-to-patch attention maps of pretrained CLIP-Base and FusionCLIP. CLIP-Base assigns substantial attention to background regions, whereas FusionCLIP masks many background patches and concentrates its attention more strongly on visible plasma structures. This qualitative difference suggests that SSS helps preserve informative local regions when forming the pretrained video representation.

Downstream Prediction Examples. Figure 4 compares predictions from FusionCLIP and CLIP-Base on the three downstream tasks. The disruption example from discharge #145327 is particularly illustrative. A visible abnormality appears around 6.734 s, while FusionCLIP issues a warning at 6.689 s, before the abnormality becomes apparent in the video. CLIP-Base does not issue a warning at that point. FusionCLIP responds again to the subsequent abnormality near 6.954 s. The ELM and H98 panels provide additional examples of the two models' downstream predictions.

6 Conclusion

We presented FusionCLIP to learn transferable representations from paired visible-light videos and diagnostic signals. SSS structures the selection of local video information, while TQE encodes heterogeneously sampled signals using their observation times. Contrastive pretraining aligns the resulting representations. On three independent diagnostic tasks in EAST-MMD, FusionCLIP achieves the highest aggregate score, and component ablations support the contributions of pretraining, SSS, and TQE. Our current evaluation is limited to EAST data. Future work will examine the cross-device transfer of this pretraining approach on other tokamaks.

AI USE STATEMENT

In this work, we used generative AI tools to assist with organizing raw data and developing program scripts for data annotation. We also used generative AI tools to review, refine, and improve code that had been independently written by the authors. All AI-assisted code was manually reviewed and verified for correctness by two authors. In addition, generative AI tools were used for language editing to improve the readability and clarity of the manuscript.

The core research ideas and overall experimental design presented in this paper were independently conceived and carried out by the authors, and generative AI tools were not used to originate these core research contributions. We carefully reviewed all content produced or modified with the assistance of generative AI and verified its correctness and appropriateness. The authors take full responsibility for the final content of this work, including all text, code, claims, and other artifacts produced or modified with the assistance of generative AI.

REFERENCES

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018. URL https://arxiv.org/abs/1803.01271.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022a. URL https://openreview.net/forum?id=p-BhZSz59o4.

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in neural information processing systems*, 35:32897–32912, 2022b.

Tobia Boschi, Andrea Loreti, Nicola C. Amorisco, Rodrigo H. Ordonez-Hurtado, Cécile Rousseau, George K. Holt, Eszter Székely, Alexander Whittle, Samuel Jackson, Adriano Agnello, Stanislas Pamela, Alessandra Pascale, Robert Akers, Juan Bernabe Moreno, Vassil Alexandrov, and Mykhaylo Zayats. TokaMind: A multi-modal transformer foundation model for tokamak plasma dynamics, 2026. URL https://arxiv.org/abs/2602.15084.

Kouroche Bouchiat, Nathaniel Chen, Peter Steiner, Azarakhsh Jalalvand, and Egemen Kolemen. Towards a foundation model for fusion: Multimodal representation learning of plasma state and control. Poster presented at the APS Division of Plasma Physics Annual Meeting, 2025. URL https://zenodo.org/records/20736607.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Caron_Emerging_Properties_in_Self-Supervised_Vision_Transformers_ICCV_2021_paper.html.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8:6085, 2018. doi: 10.1038/s41598-018-24271-9.

Junyi Chen, Longteng Guo, Jia Sun, Shuai Shao, Zehuan Yuan, Liang Lin, and Dongyu Zhang. Eve: Efficient vision-language pre-training with masked prediction and modality-aware moe. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1110–1119, 2024.

M. W. Chen, K. J. Yang, J. J. Zhu, Q. J. Wang, F. Cheng, D. L. Chen, X. Z. Gong, B. Zhang, X. H. Zhang, Y. M. Duan, B. F. Gao, Y. Wang, W. Wang, Z. F. Guan, and R. Z. Chen. Vi-DP: Low-latency video-based disruption prediction with multi-FOV fusion in EAST. *Nuclear Fusion*, 66 (1):016002, 2026a. doi: 10.1088/1741-4326/ae15a2. URL https://doi.org/10.1088/1741-4326/ae15a2.

Qiang Chen, Xiao Wang, Qingquan Yang, Hao Si, Zikang Yan, Meiwen Chen, Guosheng Xu, and Jin Tang. FusionMMT: A unified multimodal and multitask learning framework for nuclear fusion, 2026b. URL https://arxiv.org/abs/2609.26095.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9640–9649, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Chen_An_Empirical_Study_of_Training_Self-Supervised_Vision_Transformers_ICCV_2021_paper.html.

Joost Croonen, Jorge Amaya, and Giovanni Lapenta. Investigation of machine learning techniques for disruption prediction using JET data. *Plasma*, 6(1):89–102, 2023. doi: 10.3390/plasma6010008.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/0b408293619f725fd30162af057e531a-Abstract-Conference.html.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. EVA: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19358–19369, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Fang_EVA_Exploring_the_Limits_of_Masked_Visual_Representation_Learning_at_CVPR_2023_paper.html.

Xinyang Geng, Hao Liu, Lisa Lee, Dale Schuurmans, Sergey Levine, and Pieter Abbeel. Multimodal masked autoencoders learn transferable representations. *arXiv preprint arXiv:2205.14204*, 2022.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/He_Masked_Autoencoders_Are_Scalable_Vision_Learners_CVPR_2022_paper.html.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

Yuan Liu, Songyang Zhang, Jiacheng Chen, Kai Chen, and Dahua Lin. PixMIM: Rethinking pixel reconstruction in masked image modeling. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2303.02416.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Jbdc0vTOcol.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Cheng Shi, Yizhou Yu, and Sibei Yang. Vision transformers need more than registers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26328–26337, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Shi_Vision_Transformers_Need_More_Than_Registers_CVPR_2026_paper.html.

Satya Narayan Shukla and Benjamin M. Marlin. Multi-time attention networks for irregularly sampled time series. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2101.10318.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. EVA-CLIP: Improved training techniques for CLIP at scale. *arXiv preprint arXiv:2303.15389*, 2023. URL https://arxiv.org/abs/2303.15389.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19175–19186. IEEE, 2023.

Xiao Wang, Hao Si, Qiang Chen, Yu-Xiang Zhang, Beihe Zhang, Jianhua Yang, Qingquan Yang, Dengdi Sun, Wanli Lyu, Guosheng Xu, and Jin Tang. Visible-light imaging diagnosis of neutral particle emission tomography in the tokamak divertor: An efficient transformer-based surrogate model, 2026. URL https://arxiv.org/abs/2607.22704.

Xinlong Wang et al. Emu3: Next-token prediction is all you need, 2024. URL https://arxiv.org/abs/2409.18869.

Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying CLIP data. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=5BCFlnfE1g.

Zongyu Yang, Yuhang Liu, Xiaobo Zhu, Zhengwei Chen, Fan Xia, Wulyu Zhong, Zhe Gao, Yipo Zhang, and Yi Liu. Recent progress on deep learning-based disruption prediction algorithm in HL-2A tokamak. *Chinese Physics B*, 32(7):075202, 2023. doi: 10.1088/1674-1056/accb44. URL https://www.cpsjournals.cn/en/article/doi/10.1088/1674-1056/accb44.

Zongyu Yang, Zhenghao Yang, Wenjing Tian, Jiyuan Li, Xiang Sun, Guohui Zheng, Songfen Liu, Niannian Wu, Rongpeng Li, Zhaohe Xu, Bo Li, Zhongbing Shi, Zhe Gao, Wei Chen, Xiaoquan Ji, Min Xu, and Wulyu Zhong. FusionMAE, a self-supervised pretrained model to optimize and simplify diagnostic and control of fusion plasma. *Communications Physics*, 9:232, 2026. doi: 10.1038/s42005-026-02626-3. URL https://www.nature.com/articles/s42005-026-02626-3.

Wei Zheng, Fengming Xue, Zhongyong Chen, Dalong Chen, Bihao Guo, Chengshuo Shen, Xinkun Ai, Nengchao Wang, Ming Zhang, Yonghua Ding, Zhipeng Chen, Zhoujun Yang, Biao Shen, Bingjia Xiao, and Yuan Pan. Disruption prediction for future tokamaks using parameter-based transfer learning. *Communications Physics*, 6:181, 2023. doi: 10.1038/s42005-023-01296-9.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image BERT pre-training with online tokenizer. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=ydopy-e6Dg.

A Appendix

A.1 Dataset and Task Labels

EAST-MMD contains 1307 EAST discharges. 667 discharges are used for pretraining, yielding 2725021 multimodal windows of 50 ms with a stride of 2 ms. The remaining 640 discharges are used for downstream evaluation, including 481 non-disruptive and 159 disruptive discharges, and yield 280,094 labeled windows. The data are split by discharge into training, validation, and test sets containing 447, 96, and 97 discharges, respectively. This prevents windows from the same discharge from appearing in different splits.

Each window contains 12 visible-light frames and 38 diagnostic signals with their original observation timestamps. The physical meanings, units, and sampling frequencies of the signals are listed

Table 4: Time-series signals used in EAST-MMD. An asterisk denotes a diagnostic signal family that may contain multiple channels. Sampling frequencies refer to the data used in this study.

Signal	Description	Frequency (Hz)
P_{Ohm}	Ohmic heating power measurement.	100
P_{NBI}	Neutral beam injection absorbed power measurement.	100
P^*_{LHCD}	Lower hybrid current drive absorbed power measurement.	100
P^*_{ICRF}	Ion cyclotron resonance heating related power signals, describing ICRF heating conditions.	100
P^*_{ECRH}	Electron cyclotron resonance heating related power signals, describing ECRH input conditions.	100
I_p	Plasma current measurement.	10–5181
V_{loop}	Loop voltage measurement.	10–5181
z_{cur}	Vertical position of the plasma current centroid.	1–20
κ	Plasma elongation parameter describing plasma cross-section shape.	2–50
δ_{top}	Upper plasma triangularity, characterizing plasma shape and magnetic configuration.	2–50
l_i	Plasma internal inductance parameter.	2–50
q_{95}	Edge safety factor evaluated at 95% of the normalized poloidal magnetic flux, closely related to plasma MHD stability.	2–50
β_N, β_P	Normalized and poloidal plasma beta parameters, describing plasma pressure and confinement performance.	2–50
W_{MHD}	Plasma stored energy derived from MHD equilibrium reconstruction, providing an estimate of plasma energy content.	2–50
n_e	Plasma electron density measurement.	25–1000
f_{GW}	Ratio of plasma electron density to the Greenwald density limit.	1–5000
$n_{e,\mathrm{point}}$	Line-averaged electron density signal measured by the polarimeter-interferometer diagnostic, providing information on plasma density evolution.	25000
dR_{sep}	Radial distance between the primary and secondary X-point surfaces at the outer midplane.	1–20
XUV*	Extreme ultraviolet signals from different channels, providing information on plasma radiation and impurity behavior.	1000
SXR	Soft X-ray emission signal, providing information on plasma radiation.	1000
D_α	D_α emission signals from different channels, characterizing edge plasma radiation and recycling processes.	250000
NGP	Neutral gas pressure signals measured in the upper and lower divertors, providing information on divertor neutral pressure conditions.	1000
ENG	Plasma stored energy measured by the diamagnetic diagnostic.	1000–10000
SMBI	Supersonic molecular beam injection related signal.	1000
$n_{e,\mathrm{fb}}$	Line-averaged electron density signal used for density feedback control.	25–1000
$n_{e,\mathrm{target}}$	Density reference signal for feedback control in the plasma control system, specifying the target plasma density.	25–1000
I_{RMP}	Resonant magnetic perturbation coil-current signal, describing the applied magnetic perturbation.	1000
ECE*	Electron cyclotron emission signals from different channels, providing information related to electron temperature and plasma radiation.	1000
HMP	High-frequency magnetic probe signal for measuring magnetic fluctuations and characterizing MHD and Alfvénic activities.	200000

in Table 4. Figure 5 shows representative samples. Panel (a) illustrates signal evolution before a disruption and the subsequent current quench, while panels (b–c) show different ELM states. The three downstream tasks are fine-tuned and evaluated separately.

Disruption Prediction. A plasma disruption rapidly degrades confinement and can impose substantial thermal and electromagnetic loads on device components. Disruption prediction aims to identify the risk before the event, allowing time for warning and mitigation.

Each discharge is first labeled as disruptive or non-disruptive. For a disruptive discharge, let t_{dis} denote the annotated disruption time. For a millisecond-level timestamp t within a window, define the time to disruption as $\Delta t = t_{\mathrm{dis}} - t$. The soft target at t is

$$y_t = \begin{cases} 0, & \Delta t > 200, \\ \dfrac{1}{1+\exp[(\Delta t - 100)/15]}, & 10 < \Delta t \le 200, \\ 1, & \Delta t \le 10. \end{cases} \tag{13}$$

All times are measured in milliseconds. For a 50 ms window starting at t_s, the continuous window target is the mean of its 50 millisecond-level targets:

$$y_{\mathrm{dis}}(t_s) = \frac{1}{50} \sum_{t=t_s}^{t_s+49} y_t. \tag{14}$$

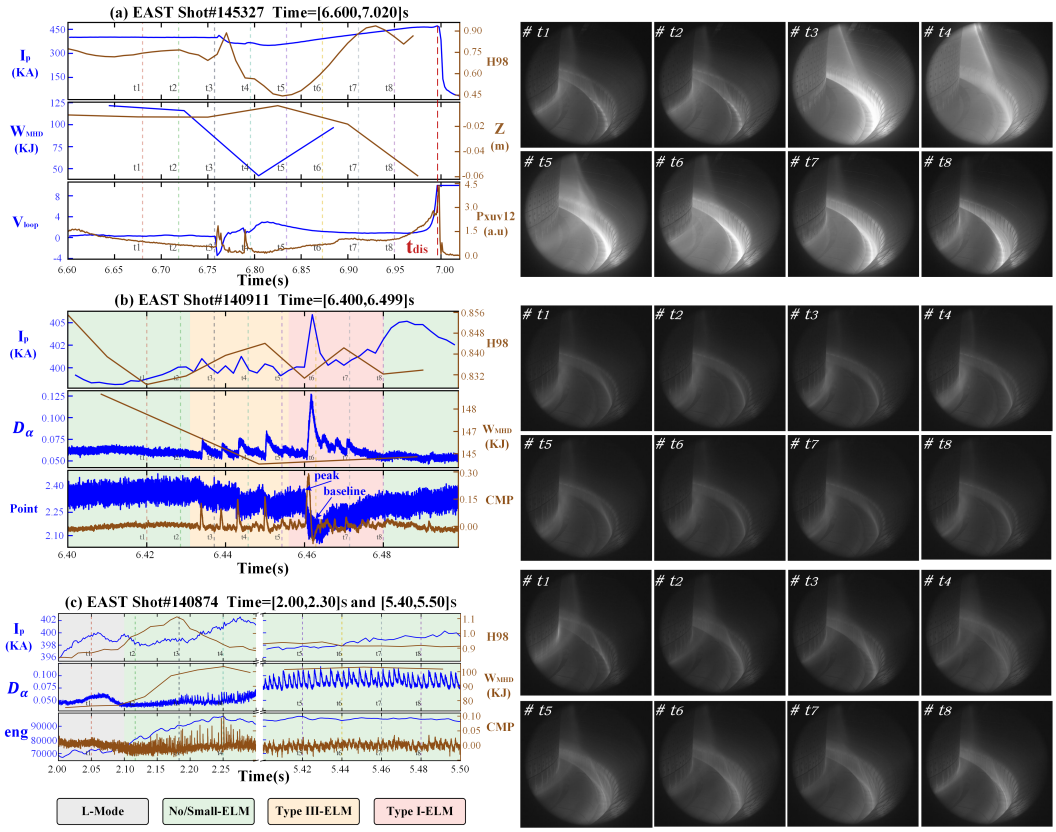

Figure 5: Illustration of representative multimodal samples in EAST-MMD. Each sample consists of synchronized time-series diagnostic signals (left) and visible-light images (right). (a) A disruptive discharge showing the pre-disruption evolution and final current quench; (b–c) time intervals with different ELM states, where the background colors indicate the corresponding ELM labels.

All windows from non-disruptive discharges have a target of zero. The disruption time is used only to generate supervision offline and is not provided as model input. During validation and testing, window predictions are processed in chronological order to produce discharge-level alarms.

ELM Classification. Edge-localized modes (ELMs) are periodic instabilities at the plasma edge during high-confinement operation. They can expel accumulated particles and impurities, but strong events may also impose transient heat loads on plasma-facing components. Distinguishing ELM states therefore helps characterize operating regimes and edge behavior.

This task uses four manually annotated operational classes: L-mode, No/Small-ELM, Type-III ELM, and Type-I ELM. Annotations are first assigned to time intervals, and each window receives the class occupying the largest portion of its duration. Examples are shown in Fig. 5(b–c).

L-mode denotes low-confinement operation and is typically associated with a lower H_{98}. During the transition to H-mode, D_α emission typically drops while stored energy and density increase. No/Small-ELM denotes intervals without a clear ELM or with only small, relatively frequent edge events. Type-III ELMs typically exhibit relatively frequent D_α bursts and modest edge-density reductions. Type-I ELMs are generally stronger and less frequent, and may be accompanied by pronounced decreases in edge density and stored energy as well as magnetic-probe responses.

Human annotation considers D_α emission, edge density, stored energy, and magnetic-probe signals together. Typical edge-density reductions are approximately 1% for small ELMs, 1%–3.5% for Type-III ELMs, and above 4% for Type-I ELMs. These values describe characteristic events and assist annotation, rather than serving as fixed thresholds for automatic label generation. The edge-

density reduction is calculated from the event peak and baseline as

$$\Delta n_{\mathrm{edge}} = \frac{\mathrm{peak} - \mathrm{baseline}}{\mathrm{peak}} \times 100\%. \tag{15}$$

H_{98} Regression. H_{98} measures the ratio of the plasma thermal-energy confinement time to the value predicted by the ITER98(y,2) scaling law. It is commonly used to assess confinement performance. This task regresses the corresponding H_{98} value from the multimodal observations within each window. The target is computed offline from diagnostic measurements and the scaling law:

$$H_{98} = \frac{\tau_{E,\mathrm{th}}}{\tau_{\mathrm{th},98y2}} = \frac{W_{\mathrm{th}}/P_{\mathrm{loss}}}{\tau_{\mathrm{th},98y2}}, \qquad P_{\mathrm{loss}} = P_{\mathrm{abs}} - \frac{\mathrm{d}W_{\mathrm{th}}}{\mathrm{d}t}. \tag{16}$$

Here, W_{th} is the thermal stored energy, P_{abs} is the absorbed heating power, and $\tau_{\mathrm{th},98y2}$ is the thermal-energy confinement time predicted by ITER98(y,2). In the available EAST processing pipeline, the EFIT-derived W_{MHD} serves as the stored-energy diagnostic for generating the offline target. The precomputed H_{98} value is not included in the model input. The model learns to predict this target from the video and diagnostic signals in the window.