SeaCausal-FL: Federated Fuzzy Causal Learning for Maritime IoT Fault Diagnosis and Counterfactual Reasoning

Yuhang Qiu, Haihan Zhu, Koteeswaran Seerangan, Longsheng Zhu, Xiong Wang, Yijun Lu, Zheng Lin, Fangmin Ren and Jialiang Xie

***Abstract*—Reliable marine-engine fault diagnosis in maritime IoT is challenged by distributed data ownership, heterogeneous fault distributions, and continuously changing operating conditions. This paper proposes SeaCausal-FL, a federated fuzzy causal learning framework that combines a shared temporal diagnostic path with mechanism-conditioned causal reasoning. An interval type-2 fuzzy layer represents uncertain and overlapping operating mechanisms, while each mechanism is associated with a physics-constrained structural causal model. Before aggregation, locally learned mechanisms are aligned using operating context, causal structure, and conditional intervention–response signatures. Model parameters are then aggregated according to sample, class, mechanism, and mechanism–class evidence instead of client sample size alone. The learned structural equations further support interval counterfactual reasoning through abduction, action, and prediction. Experiments on a marine-engine fault dataset and a real-data-calibrated semi-synthetic causal benchmark show that SeaCausal-FL achieves an average F1-score of 87.07% across four client partitions, with AUROC and AUPRC of 98.98% and 94.81%, respectively. It also maintains strong performance under unseen loads and fault-type omission during training. On the causal benchmark, SeaCausal-FL reaches an Edge-F1 of approximately 0.58 and an Edge-AUPRC of 0.68, reduces coefficient RMSE to about 0.14, and provides favorable counterfactual estimation and intervention decisions.**

***Index Terms*—Maritime IoT fault diagnosis, Causal reasoning, federated learning, interval type-2 fuzzy systems.**

Yuhang Qiu and Haihan Zhu contributed equally to this work. This work was supported in part by the National Natural Science Foundation of China (No. 12571585) and Scientific Research Start-up Program of Jimei University (No. Z8268125).

Yuhang Qiu, Fangmin Ren, and Jialiang Xie are with the School of Science, Jimei University, Xiamen 361021, China. (e-mail: yuhang.qiu@jmu.edu.cn; fangminren@jmu.edu.cn; xiejialiang@jmu.edu.cn).

Haihan Zhu is with the School of Artificial Intelligence, Hebei University of Technology, Tianjin 300401, China (email: 242794@stu.hebut.edu.cn)

Koteeswaran Seerangan is with the Department of Computer Science and Engineering, R. M. K. Engineering College (Autonomous), Kavaraipettai - 601206, Tamil Nadu, India (e-mail: skn.cse@rmkec.ac.in).

Longsheng Zhu is with the College of Information Network Security, People's Public Security University of China, Beijing 100038, China (e-mail: 2024111028@stu.ppsuc.edu.cn).

Xiong Wang is with the School of Computer Science and Technology, University of Science and Technology of China, Hefei 230026, China (e-mail: wangxiong@ustc.edu.cn).

Yijun Lu is with the Department of Computer Science and Engineering, Waseda University, Tokyo, Japan (e-mail: yijun@ruri.waseda.jp).

Zheng Lin is with the Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg, Luxembourg (e-mail: zhenglin@ieee.org).

(Corresponding author: Jialiang Xie)

I. Introduction

MARINE transportation is increasingly dependent on onboard sensing, edge computing, and shore-side intelligent services. Under the Internet of Ships and maritime Internet of Things paradigms, modern vessels routinely collect multivariate signals related to engine load, speed, fuel supply, cooling, combustion, exhaust conditions, and power output [1–3]. These measurements provide an important basis for marine-engine fault diagnosis by supporting the early detection of abnormal operating states before they develop into propulsion failures, which is important for navigation safety, maintenance cost reduction, and vessel availability [4–6].

Machine learning, as a powerful data-driven paradigm for extracting discriminative representations [7–11] has been widely applied to maritime component monitoring and marine diesel-engine diagnosis [5, 6, 12, 13]. However, diagnostic performance remains sensitive to operating-condition variation [12–14]. The same fault can produce different sensor responses under different loads or thermal states, while different faults can cause similar changes in pressure, temperature, or flow. Moreover, operating conditions commonly vary continuously rather than forming clearly separated regimes. These characteristics create heterogeneous and overlapping fault patterns that are difficult to represent with a single condition-independent diagnostic model.

A further challenge arises from the distributed ownership of maritime data. Measurements collected from different vessels, engine units, maintenance organizations, or test campaigns are generally stored at separate sites, while direct data sharing can be restricted by commercial confidentiality, communication cost, and data-governance requirements. Federated learning provides a practical way to train diagnostic models without transferring raw measurements [15–17]. Existing federated fault-diagnosis studies have addressed data heterogeneity and domain shifts through transfer learning, domain generalization, meta-learning, and personalized aggregation [18–24]. Hierarchical split federated learning has also been studied for coordinating model splitting and aggregation across multi-tier edge systems, further illustrating the importance of adapting distributed learning to heterogeneous computing environments [25]. These methods can improve diagnostic robustness under several forms of data heterogeneity, but their coordination units are generally complete models, clients, domains, latent features, or fault-class representations.

Such coordination does not explicitly distinguish multiple uncertain operating mechanisms that can coexist within the same client. It also does not guarantee that locally learned components with the same parameter index represent physically corresponding operating mechanisms across clients. As a result, direct aggregation can mix operating-specific knowledge with different physical meanings. Class-aware aggregation alleviates label skew but does not determine whether the available class evidence is supported under the same operating mechanism. In addition, most federated diagnostic models are optimized primarily for fault recognition and do not preserve mechanism-specific physical relations required for intervention and counterfactual analysis. These limitations motivate a finer-grained framework that coordinates operating mechanisms rather than only complete client models or class-level representations.

To address these limitations, this paper proposes SeaCausal-FL, a federated fuzzy causal learning framework for maritime IoT fault diagnosis and counterfactual reasoning. SeaCausal-FL combines a shared temporal diagnostic path with an operating-dependent fuzzy causal path. An interval type-2 fuzzy layer [26, 27] represents uncertain and overlapping operating mechanisms through lower, midpoint, and upper membership weights. Each mechanism parameterizes a structural causal model within a physics-constrained candidate graph [28, 29], while its mechanism-specific output acts as a residual correction to the shared diagnosis. The residual contribution is initialized at zero so that the common diagnostic representation is established before mechanism-specific causal corrections become active.

Before aggregation, locally learned mechanisms are aligned with the broadcast global reference using operating context and causal structure. When the preliminary alignment indicates a possible permutation, intervention–response signatures are further used to refine the correspondence. Different parameter groups are then aggregated according to their supporting evidence: shared parameters use sample counts, class-specific parameters use class counts, and mechanism-related parameters use fuzzy membership and mechanism–class evidence. The learned structural equations further support counterfactual inference through abduction, action, and prediction, providing nominal and interval-valued fault-risk estimates under feasible physical interventions.

The main contributions of this paper are summarized as follows.

1) We propose a federated diagnostic architecture for simultaneous operating-condition and label heterogeneity. It combines a shared temporal diagnostic model with mechanism-conditioned causal residuals to preserve stable fault recognition while modeling operating-dependent physical responses.
2) We introduce an interval type-2 fuzzy causal mechanism model for uncertain and overlapping operating conditions. Each fuzzy mechanism parameterizes a structural causal model within a physics-constrained candidate graph, while its center, uncertainty footprint, edge evidence, and structural coefficients are learned from training data rather than manually defined operating thresholds or fixed causal weights.
3) We develop a physical-response-based mechanism alignment and mechanism class-aware aggregation strategy. Local mechanisms are matched according to operating context, causal structure, and intervention response, and their parameters are aggregated using effective fuzzy membership mass and class evidence rather than client sample size alone.
4) We evaluate SeaCausal-FL on a real marine-engine fault dataset and a real-data-calibrated semi-synthetic causal benchmark. The experiments cover natural non-IID, IID, and Dirichlet partitions, unseen operating loads, incomplete fault-type coverage, client participation, causal structure recovery, and counterfactual estimation.

The remainder of this paper is organized as follows. Section II reviews related work. Section III presents the proposed SeaCausal-FL framework. Section IV describes the experimental settings and reports the results. Section V concludes the paper.

II. Related Work

Marine engine fault diagnosis has gradually shifted from expert rules and conventional signal analysis to machine learning and deep learning methods [30–33]. Existing studies have used ensemble learning, manifold learning, recurrent networks, and transfer learning to extract diagnostic information from pressure, temperature, vibration, and other engine measurements [5, 6, 12]. Recent cross-condition methods further employ domain adaptation, feature disentanglement, and centroid alignment to reduce distribution differences among loads, machines, and operating environments [13, 14]. These methods improve recognition under changing conditions, but they generally regard each load or machine as a predefined domain or seek representations that remove operating-condition information. Such treatment is less suitable when operating regimes change gradually, overlap with one another, or produce genuinely different physical responses to the same fault. Most of these methods also assume centralized access to the diagnostic data.

Federated fault diagnosis has received increasing attention. Early studies combined self-supervised local training with validation-guided aggregation, while federated transfer-learning methods used prior distributions, feature adaptation, and target self-adaptation to handle cross-machine and cross-condition shifts [18, 19, 34]. Federated domain generalization and meta-learning have also been studied for unseen domains, limited local samples, and few-shot diagnosis [20, 35, 36]. More recent methods use class information, personalized classifiers, and fault prototypes to address label-distribution skew [21, 37]. Personalized aggregation has further been considered when a client contains several working conditions, while communication-aware and open-set methods have been developed for offshore equipment and unknown faults [22, 23, 38]. These studies cover several important forms of data heterogeneity, but their coordination units remain complete models, clients, domains, latent features, or fault-class prototypes. They do not explicitly determine whether local components learned at different clients describe the same operating mechanism.

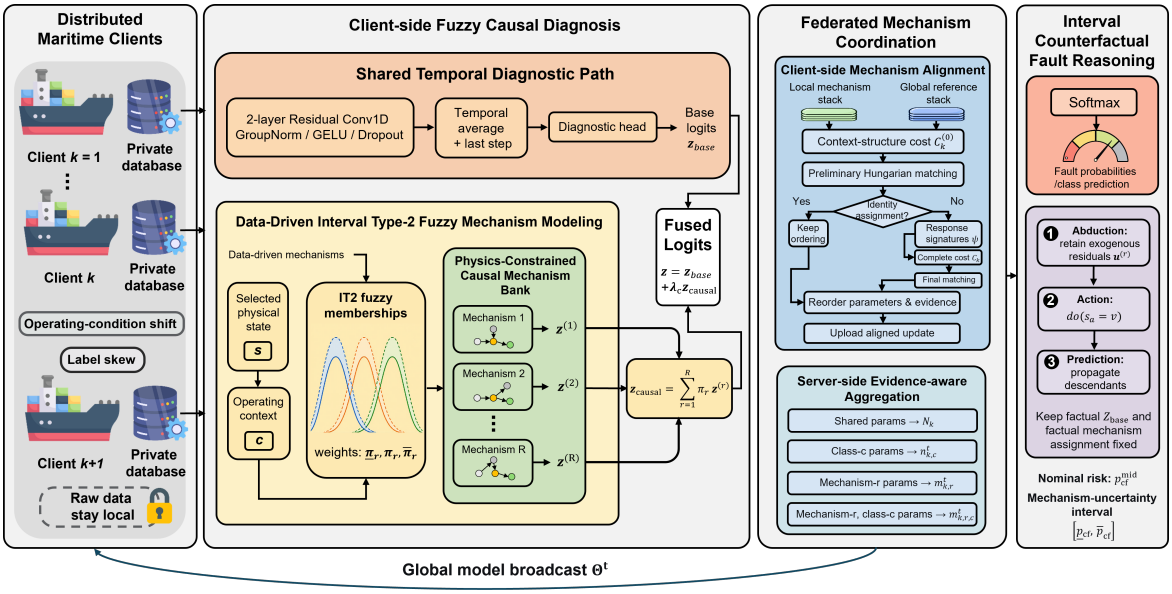

Fig. 1. Overall framework of SeaCausal-FL. Each maritime client performs client-side fuzzy causal diagnosis through a shared temporal diagnostic path and an operating-dependent fuzzy causal path. Locally updated mechanisms are aligned with the broadcast global reference before upload. The server then performs evidence-aware aggregation using sample, class, mechanism, and mechanism–class evidence. The resulting global model supports both fault diagnosis and interval counterfactual reasoning.

Class-aware aggregation identifies available fault evidence, but it does not determine whether that evidence is obtained under a corresponding physical mechanism. As a result, operating-specific knowledge with different meanings can still be mixed during aggregation.

Fuzzy systems provide a natural way to describe operating regimes without imposing hard boundaries. In particular, interval type-2 fuzzy models represent uncertainty in the membership itself and have been applied to nonstationary industrial processes [26, 27]. In existing diagnostic models, however, fuzzy memberships are mainly used for rule activation, sample weighting, feature mapping, or prediction confidence. They do not usually define distinct physical mechanisms. Structural causal models offer a complementary means of representing how changes in one variable propagate to downstream variables. Recent federated causal-learning methods recover global causal graphs from decentralized and heterogeneous data, with later studies considering scalability and nonidentical variable sets across clients [39–41]. Federated causal representations have also begun to support counterfactual reasoning in distributed industrial systems [42]. However, these methods mainly seek a single global graph or a shared latent causal representation and are not designed for fault diagnosis with overlapping operating mechanisms and incomplete local fault classes. SeaCausal-FL connects these directions by associating each uncertain operating mechanism with a local structural causal model, aligning corresponding mechanisms before aggregation, and conditioning class-specific parameter sharing on both mechanism activation and local fault evidence.

III. The SeaCausal-FL Framework

A. System Model and Learning Objective

As shown in Fig. 1, SeaCausal-FL considers a federated maritime IoT system consisting of distributed maritime clients and a federated server. The main components are described as follows.

- **Maritime clients:** Let $\mathcal{K} = \{1, 2, \ldots, K\}$ denote the set of participating clients. Each client stores a private collection of marine-engine sensor measurements and performs model training locally.
- **Local sensor data:** The continuous sensor sequence at client k is divided into fixed-length sliding windows. The local training dataset is denoted by

$$\mathcal{D}_k = \{(\mathbf{X}_{k,n}, \mathbf{M}_{k,n}, y_{k,n})\}_{n=1}^{N_k}, \tag{1}$$

  where N_k is the number of valid training windows at client k, $\mathbf{X}_{k,n} \in \mathbb{R}^{T\times D}$ is the nth sensor window with T time steps and D retained variables, and $\mathbf{M}_{k,n} \in \{0,1\}^{T\times D}$ is the corresponding missing-value mask, in which an entry of one indicates a missing observation. The label $y_{k,n} \in \mathcal{C} = \{0, \ldots, C-1\}$ identifies the fault class. A window is retained only when all of its observations belong to the same temporal partition and fault state, preventing temporal leakage and mixed-label supervision.
- **Federated server:** The server initializes and broadcasts the global model, receives the aligned local parameters and compact evidence statistics, and performs evidence-aware model aggregation.

Let $P_k(\mathbf{X}, \mathbf{M}, Y)$ denote the local data distribution at client k. SeaCausal-FL does not impose a specific client-partition assumption. Under an IID partition, $P_k(\mathbf{X}, \mathbf{M}, Y) = P_j(\mathbf{X}, \mathbf{M}, Y)$ for all client pairs. Under a heterogeneous partition, there exists at least one pair of clients satisfying $P_k(\mathbf{X}, \mathbf{M}, Y) \neq P_j(\mathbf{X}, \mathbf{M}, Y)$ for $k \neq j$. Such heterogeneity is induced by operating-condition variation, label skew, or both. The same fault can produce different sensor responses

under different loads, while some clients contain only a subset of the fault classes. In these cases, conventional sample-count-based aggregation does not distinguish how much evidence a client provides for a particular operating mechanism or fault class.

Let $\mathbf{\Theta}$ denote all trainable parameters of SeaCausal-FL, and let $\mathbf{p}_{k,n}(\mathbf{\Theta}) \in [0,1]^C$ denote the predicted fault-probability vector for the nth window of client k. The local learning objective is defined as

$$\mathcal{F}_k(\mathbf{\Theta}) = \frac{1}{N_k} \sum_{n=1}^{N_k} \ell_{\mathrm{cw}} \left(\mathbf{p}_{k,n}(\mathbf{\Theta}), y_{k,n} \right) + \mathcal{L}_k^{\mathrm{aux}}(\mathbf{\Theta}), \tag{2}$$

where $\ell_{\mathrm{cw}}(\cdot)$ denotes the class-weighted cross-entropy loss. The auxiliary term $\mathcal{L}_k^{\mathrm{aux}}$ contains the objectives used for causal-branch supervision, fuzzy-mechanism regularization, structural-equation reconstruction, and causal-edge regularization. These objectives are introduced together with their corresponding modules in the following subsections.

Let $N = \sum_{k=1}^{K} N_k$ denote the total number of training windows across all clients. A reference sample-weighted federated objective is formulated as

$$\min_{\mathbf{\Theta}} \mathcal{F}(\mathbf{\Theta}) = \sum_{k=1}^{K} \frac{N_k}{N} \mathcal{F}_k(\mathbf{\Theta}). \tag{3}$$

Equation (3) describes the conventional client-level empirical weighting. SeaCausal-FL retains the coefficient N_k/N only for shared non-class parameters rather than applying it uniformly to all parameter groups. Class-specific, mechanism-specific, and mechanism–class-specific parameters are aggregated according to local class counts, effective mechanism membership masses, and joint mechanism–class evidence, respectively.

B. Overview of SeaCausal-FL

Fig. 1 presents the overall architecture of SeaCausal-FL. At communication round t, the server broadcasts the current global model to a subset of maritime clients. Each selected client updates the model using its private sensor windows, aligns its locally learned mechanisms with the broadcast global reference, and uploads the aligned parameters together with compact evidence statistics for server aggregation. The framework contains three main components.

1) Client-Side Fuzzy Causal Diagnosis: Each client first uses a shared temporal encoder to process the sensor window and its missing-value mask. The resulting representation is mapped to the base diagnostic logits as

$$\mathbf{z}_{k,n}^{\mathrm{base}} = g_{\boldsymbol{\theta}_d} \left(f_{\boldsymbol{\theta}_e} \left(\mathbf{X}_{k,n}, \mathbf{M}_{k,n} \right) \right), \tag{4}$$

where $f_{\boldsymbol{\theta}_e}(\cdot)$ and $g_{\boldsymbol{\theta}_d}(\cdot)$ denote the shared temporal encoder and diagnostic head, respectively. This shared path learns fault characteristics that remain useful across different operating conditions.

To model operating-dependent physical responses, SeaCausal-FL extracts a physical state $\mathbf{s}_{k,n}$ from selected engine variables. Engine speed, water-brake load, and engine-room temperature form the operating context $\mathbf{c}_{k,n}$. An interval type-2 fuzzy layer assigns lower, midpoint, and upper weights, denoted by $\underline{\pi}_{k,n,r}$, $\pi_{k,n,r}$, and $\overline{\pi}_{k,n,r}$, respectively, to each latent operating mechanism $r \in \{1, \ldots, R\}$. The number of mechanisms is selected from the training operating contexts subject to a data-sufficiency constraint. The fuzzy centers, lower and upper scales, and mixture priors are initialized from the resulting data-driven prior and further optimized during local training.

Each operating mechanism is associated with a structural causal model constructed within a physics-constrained candidate graph. Let $\mathbf{z}_{k,n}^{(r)} \in \mathbb{R}^C$ denote the mechanism-specific causal logits produced by mechanism r. The midpoint fuzzy weights combine these mechanism-specific outputs as

$$\mathbf{z}_{k,n}^{\mathrm{causal}} = \sum_{r=1}^{R} \pi_{k,n,r} \mathbf{z}_{k,n}^{(r)}. \tag{5}$$

The weighted causal output is introduced as a residual correction to the shared diagnostic logits:

$$\mathbf{z}_{k,n} = \mathbf{z}_{k,n}^{\mathrm{base}} + \lambda_{\mathrm{c}} \mathbf{z}_{k,n}^{\mathrm{causal}}, \tag{6}$$

where λ_{c} is a bounded learnable coefficient obtained from ρ through a hyperbolic-tangent transformation. Since ρ is initialized to zero, the shared diagnostic path determines the initial prediction, while the causal correction is introduced gradually during training.

2) Federated Mechanism Coordination: Local mechanism indices do not necessarily share the same physical meaning across clients. Before upload, each selected client aligns its locally updated mechanisms with the broadcast global reference using operating-context centers and causal structures. If the preliminary assignment is nonidentity, label-independent intervention–response signatures are added to refine the matching. The aligned mechanism parameters and evidence statistics are then reordered before transmission.

The server applies evidence-aware aggregation to different parameter groups. Shared parameters use local sample counts, class-specific parameters use class counts, mechanism-specific fuzzy and causal parameters use effective membership masses, and mechanism–class-specific parameters use joint mechanism–class evidence. Clients with insufficient support therefore exert limited influence on the corresponding global parameters.

3) Interval Counterfactual Fault Reasoning: The global model performs fault diagnosis using $\mathbf{p}_{k,n} = \mathrm{softmax}(\mathbf{z}_{k,n})$ and supports interval counterfactual reasoning through abduction, action, and prediction. Given a factual physical state, SeaCausal-FL retains the mechanism-specific exogenous residuals, intervenes on an actionable variable, and propagates its effects through the downstream causal nodes. The factual temporal logits and mechanism assignment remain fixed. The midpoint mechanism weights produce the nominal counterfactual risk, while the feasible IT2 weight set defined by the lower and upper bounds determines its interval.

The complete training procedure is summarized in Algorithm 1. In each communication round, selected clients jointly train the shared diagnostic and fuzzy causal components, collect mechanism-related evidence, and align their

Algorithm 1 Training Procedure of SeaCausal-FL

Input: Local datasets $\{\mathcal{D}_k\}_{k=1}^{K}$; data-driven prior $\mathcal{P}$; intervention grid $\mathcal{I}$; communication rounds T_{FL}; local epochs E; participation ratio q; class weights $\boldsymbol{\omega}$; patience H.
Output: Best global model $\boldsymbol{\Theta}^{\star}$.

Initialize $\boldsymbol{\Theta}^{0}$ from $\mathcal{P}$; set $s^{\star} \leftarrow -\infty$ and $p \leftarrow 0$.
for $t = 0, 1, \ldots, T_{\mathrm{FL}} - 1$ **do**
  Select $\mathcal{S}_t \subseteq \mathcal{K}$ according to q and broadcast $\boldsymbol{\Theta}^{t}$.
  Client-side learning and mechanism alignment
  for all $k \in \mathcal{S}_t$ **in parallel do**
    Set $\boldsymbol{\Theta}_k \leftarrow \boldsymbol{\Theta}^{t}$ and initialize $m^{t}_{k,r}$, $m^{t}_{k,r,c}$, and $n^{t}_{k,c}$ to zero.
    for $e = 1, 2, \ldots, E$ **do**
      for all minibatches $\mathcal{B} \subset \mathcal{D}_k$ **do**
        Compute the shared, fuzzy, causal, and fused outputs.
        Compute $\mathcal{L}_k$ using (30) and update $\boldsymbol{\Theta}_k$ using AdamW with gradient clipping.
        Accumulate $m^{t}_{k,r}$, $m^{t}_{k,r,c}$, and $n^{t}_{k,c}$ according to (41).
      end for
    end for
    Set $\boldsymbol{\Theta}^{t+1}_{k} \leftarrow \boldsymbol{\Theta}_k$ and collect $\mathcal{E}^{t}_{k} \leftarrow \{N_k, m^{t}_{k,r}, m^{t}_{k,r,c}, n^{t}_{k,c}\}$.
    Construct $\mathbf{C}^{(0)}_{k}$ using (34) and obtain $\sigma^{(0)}_{k}$ by Hungarian matching.
    if $\sigma^{(0)}_{k}$ is not the identity assignment **then**
      Compute local and reference intervention–response signatures on the same private calibration minibatches.
      Construct $\mathbf{C}_k$ using (39) and obtain σ^{t}_{k} by Hungarian matching.
    else
      Set $\sigma^{t}_{k} \leftarrow \sigma^{(0)}_{k}$.
    end if
    Jointly reorder $\boldsymbol{\Theta}^{t+1}_{k}$ and $\mathcal{E}^{t}_{k}$ according to σ^{t}_{k}.
    Upload the aligned local parameters and evidence statistics.
  end for
  Evidence-aware server aggregation
  Aggregate the four parameter groups using N_k, $n^{t}_{k,c}$, $m^{t}_{k,r}$, and $m^{t}_{k,r,c}$ according to (42).
  Retain the previous global parameter block when its total evidence is zero.
  Obtain $\boldsymbol{\Theta}^{t+1}$ and its validation F1-score s^{t+1}.
  if $s^{t+1} > s^{\star}$ **then**
    $\boldsymbol{\Theta}^{\star} \leftarrow \boldsymbol{\Theta}^{t+1}$, $s^{\star} \leftarrow s^{t+1}$, and $p \leftarrow 0$.
  else
    $p \leftarrow p + 1$.
  end if
  if $p \geq H$ **then**
    break
  end if
end for
return $\boldsymbol{\Theta}^{\star}$.

local mechanisms before upload. The server then aggregates different parameter groups using the corresponding sample, class, mechanism, and mechanism–class evidence.

C. Client-Side Fuzzy Causal Diagnosis

At each selected client, SeaCausal-FL jointly learns a shared temporal diagnostic path and an operating-dependent fuzzy causal path. The shared path captures fault characteristics that recur across clients, while the causal path models physical relations that vary with operating conditions. Instead of training an independent classifier for every mechanism, SeaCausal-FL retains a shared diagnostic backbone and introduces lightweight mechanism-conditioned causal corrections.

1) Shared Temporal Diagnostic Path: The shared path provides a stable diagnostic basis before the operating-dependent causal correction becomes active. Since missing observations may affect the reliability of temporal features, the sensor window and its missing-value mask are concatenated along the feature dimension. The resulting tensor is processed by a two-layer one-dimensional residual convolutional encoder. Each convolutional layer is followed by group normalization, while GELU activation and dropout are applied within the residual path. Unlike batch normalization, group normalization does not rely on client-specific batch statistics and is therefore suitable for heterogeneous federated clients. The detailed architecture of this shared temporal diagnostic path is illustrated in Fig. 2.

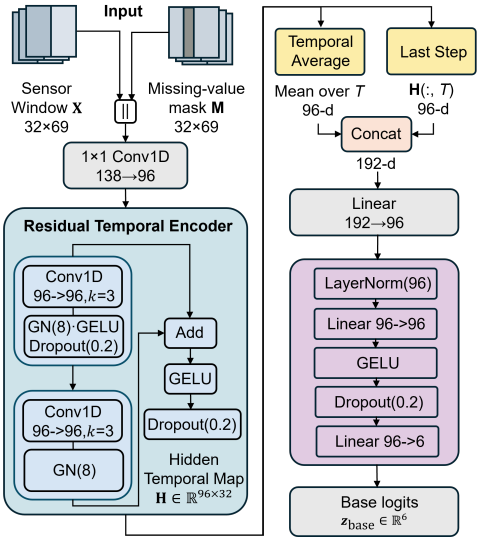

Fig. 2. Detailed architecture of the shared temporal diagnostic path.

Let $\mathbf{H}_{k,n} \in \mathbb{R}^{d_h \times T}$ denote the hidden temporal feature map produced from $(\mathbf{X}_{k,n}, \mathbf{M}_{k,n})$, where d_h is the hidden-channel dimension. SeaCausal-FL combines the average temporal response and the last-step feature as

$$\mathbf{h}_{k,n} = \mathbf{W}_p \left[\mathrm{Avg}_{\tau} \left(\mathbf{H}_{k,n} \right) \,\|\, \mathbf{H}_{k,n}(:, T) \right] + \mathbf{b}_p, \tag{7}$$

where $\|$ denotes vector concatenation. The average summary describes the overall behavior of the sensor window, while the last-step feature retains its most recent temporal response. The representation $\mathbf{h}_{k,n}$ is then mapped to the base diagnostic logits through the diagnostic head defined in (4). This shared path uses all retained sensor variables and provides the primary fault-classification signal during local training.

2) Data-Driven Interval Type-2 Fuzzy Mechanism Modeling: Fixed low-, medium-, and high-load thresholds are inappropriate when maritime operating conditions vary continuously and overlap. SeaCausal-FL therefore derives latent operating mechanisms from training data and represents uncertainty in their assignments using interval type-2 fuzzy memberships.

For each physical variable, the most recent valid value in the current sensor window is retained to construct the physical state $\mathbf{s}_{k,n}$. If a physical variable has no valid observation in the window, its value is set to zero and its own structural reconstruction term is masked out. Engine speed, water-brake load, and engine-room temperature are selected as exogenous

operating variables, giving

$$\mathbf{c}_{k,n} = \mathbf{P}_c \mathbf{s}_{k,n}, \tag{8}$$

where $\mathbf{P}_c$ is a fixed operating-context selection matrix. The membership function depends only on these exogenous operating variables; fault labels and downstream physical responses are not included in $\mathbf{c}_{k,n}$.

To estimate the data-driven prior, a diagonal-covariance Gaussian mixture model (GMM) is fitted to the operating contexts extracted from normal training windows when the number of normal windows satisfies the minimum prior-estimation requirement $N_{\text{prior}}^{\min}$; otherwise, all training contexts are used. Candidate mechanism numbers are considered within $\mathcal{R} = \{1, \ldots, R_{\max}\}$.

For an R-component GMM, let $\gamma_{n,r}^{(R)}$ denote the posterior responsibility of mechanism r for training context n. Its soft effective sample count is defined as

$$N_r^{\text{soft}}(R) = \sum_n \gamma_{n,r}^{(R)}. \tag{9}$$

The corresponding hard-assignment count is

$$N_r^{\text{hard}}(R) = \sum_n \mathbb{I}\left\{ r = \arg\max_j \gamma_{n,j}^{(R)} \right\}. \tag{10}$$

A candidate R is considered eligible only when every mechanism has sufficient support. Specifically,

$$\min_r N_r^{\text{soft}}(R) \geq N_{\text{soft}}^{\min}, \tag{11}$$

and

$$\min_r N_r^{\text{hard}}(R) \geq N_{\text{hard}}^{\min}. \tag{12}$$

The candidates satisfying both conditions form the eligible set $\mathcal{R}_{\text{elig}}$.

Among these candidates, the mechanism number is selected by

$$R^\star = \arg\min_{R \in \mathcal{R}_{\text{elig}}} \left(\text{BIC}_R + 2\mathcal{H}_R \right), \tag{13}$$

where the posterior assignment entropy is

$$\mathcal{H}_R = -\sum_n \sum_{r=1}^{R} \gamma_{n,r}^{(R)} \log \gamma_{n,r}^{(R)}. \tag{14}$$

The BIC term favors a parsimonious fit to the operating-context distribution, while the entropy term penalizes ambiguous mechanism assignments. The number of fuzzy causal mechanisms used in subsequent training is set to $R = R^\star$.

The resulting GMM centers, scales, and mixture probabilities initialize the trainable fuzzy parameters. The lower and upper scales account for finite-sample uncertainty in both scale and center estimation rather than using a manually chosen footprint width. Let $\boldsymbol{\xi}_r$ denote the center of mechanism r. For $a \in \{\text{L}, \text{U}\}$, the membership candidate obtained from scale $\boldsymbol{\sigma}_r^a$ is

$$\widetilde{\mu}_{k,n,r}^a = \exp\left[-\frac{1}{2} \left\| \frac{\mathbf{c}_{k,n} - \boldsymbol{\xi}_r}{\boldsymbol{\sigma}_r^a} \right\|_2^2 \right]. \tag{15}$$

The lower and upper memberships are defined as $\underline{\mu}_{k,n,r} = \min\{\widetilde{\mu}_{k,n,r}^{\text{L}}, \widetilde{\mu}_{k,n,r}^{\text{U}}\}$ and $\overline{\mu}_{k,n,r} = \max\{\widetilde{\mu}_{k,n,r}^{\text{L}}, \widetilde{\mu}_{k,n,r}^{\text{U}}\}$, respectively.

Let α_r denote the trainable mixture probability initialized from the GMM, with $\alpha_r \geq 0$ and $\sum_{r=1}^R \alpha_r = 1$. The midpoint mechanism weight is

$$\pi_{k,n,r} = \frac{\alpha_r \left(\underline{\mu}_{k,n,r} + \overline{\mu}_{k,n,r} \right)}{\sum_{j=1}^{R} \alpha_j \left(\underline{\mu}_{k,n,j} + \overline{\mu}_{k,n,j} \right) + \varepsilon}. \tag{16}$$

The interval memberships define a normalized uncertainty half-width around the midpoint mechanism weight. For mechanism r, it is computed as

$$\delta_{k,n,r} = \frac{\frac{1}{2}\alpha_r \left(\overline{\mu}_{k,n,r} - \underline{\mu}_{k,n,r} \right)}{\frac{1}{2}\sum_{j=1}^{R} \alpha_j \left(\underline{\mu}_{k,n,j} + \overline{\mu}_{k,n,j} \right) + \varepsilon}. \tag{17}$$

Collecting these values gives $\boldsymbol{\delta}_{k,n} = [\delta_{k,n,1}, \ldots, \delta_{k,n,R}]^\top$. The provisional lower and upper bounds are then constructed separately as

$$\mathbf{l}_{k,n} = \left[\boldsymbol{\pi}_{k,n} - \boldsymbol{\delta}_{k,n} \right]_{[0,1]}, \tag{18}$$

and

$$\mathbf{u}_{k,n} = \left[\boldsymbol{\pi}_{k,n} + \boldsymbol{\delta}_{k,n} \right]_{[0,1]}, \tag{19}$$

where $[\cdot]_{[0,1]}$ denotes element-wise clipping. If $\mathbf{1}^\top \mathbf{l}_{k,n} > 1$, the lower bound is normalized by its sum. If $\mathbf{1}^\top \mathbf{u}_{k,n} < 1$, the missing probability mass is distributed over the remaining upper-bound capacity. The bounds are finally adjusted to enclose the midpoint weight, yielding

$$\underline{\boldsymbol{\pi}}_{k,n} \preceq \boldsymbol{\pi}_{k,n} \preceq \overline{\boldsymbol{\pi}}_{k,n}, \tag{20}$$

with

$$\mathbf{1}^\top \underline{\boldsymbol{\pi}}_{k,n} \leq 1 \leq \mathbf{1}^\top \overline{\boldsymbol{\pi}}_{k,n}. \tag{21}$$

Thus, the interval contains at least one valid normalized mechanism-weight vector.

During local training, three auxiliary terms constrain the fuzzy mechanism model. Let $\boldsymbol{\xi}_r^0$, $\boldsymbol{\sigma}_r^{L,0}$, $\boldsymbol{\sigma}_r^{U,0}$, and α_r^0 denote the GMM-derived prior parameters. The prior uncertainty scale is defined as $\boldsymbol{\tau}_r = (\boldsymbol{\sigma}_r^{U,0} - \boldsymbol{\sigma}_r^{L,0})/2$. The context negative log-likelihood is

$$\mathcal{L}_k^{\text{context}} = -\frac{1}{|\mathcal{B}|} \sum_{n \in \mathcal{B}} \log \sum_{r=1}^{R} \frac{\alpha_r}{\prod_d \sigma_{r,d}^L} \exp\left[-\frac{1}{2} \left\| \frac{\mathbf{c}_{k,n} - \boldsymbol{\xi}_r}{\boldsymbol{\sigma}_r^L} \right\|_2^2 \right] \tag{22}$$

The normalized prior-drift loss is

$$\mathcal{L}_k^{\text{drift}} = \text{Mean}\left[\left(\frac{\boldsymbol{\xi} - \boldsymbol{\xi}^0}{\boldsymbol{\tau}} \right)^2 + \left(\frac{\boldsymbol{\sigma}^L - \boldsymbol{\sigma}^{L,0}}{\boldsymbol{\tau}} \right)^2 + \left(\frac{\boldsymbol{\sigma}^U - \boldsymbol{\sigma}^{U,0}}{\boldsymbol{\tau}} \right)^2 \right] \tag{23}$$

Deviations of the trainable mixture probabilities from the GMM prior are controlled by

$$\mathcal{L}_k^{\text{mix}} = \sum_{r=1}^{R} \alpha_r \log \frac{\alpha_r}{\alpha_r^0}. \tag{24}$$

These terms permit local adaptation while constraining the fuzzy mechanisms to remain close to their data-derived operating meanings.

3) Physics-Constrained Causal Mechanism Bank: Learning an unrestricted causal graph from a small and non-IID local dataset may produce cyclic or physically implausible relations. SeaCausal-FL therefore defines an auditable candidate graph according to the ordered engine processes of control, air intake, cooling, combustion, exhaust, and shaft power. The graph specifies admissible causal directions and expected coefficient signs, while edge probabilities and coefficient magnitudes remain trainable.

Let $\mathcal{E}_{\mathrm{phy}}$ denote the physics-constrained candidate edge set, and let $\mathrm{Pa}(j)$ denote the candidate parents of physical node j. For mechanism r, the edge probability associated with $(i,j) \in \mathcal{E}_{\mathrm{phy}}$ is $g_{ij}^{(r)} = \mathrm{sigmoid}(\eta_{ij}^{(r)})$. The corresponding sign-constrained linear and saturating coefficients are denoted by $a_{ij}^{(r)}$ and $c_{ij}^{(r)}$, respectively. The structural prediction of node j is

$$\widehat{s}_{k,n,j}^{(r)} = b_j^{(r)} + \sum_{i \in \mathrm{Pa}(j)} g_{ij}^{(r)} \left[a_{ij}^{(r)} s_{k,n,i} + c_{ij}^{(r)} \tanh\left(s_{k,n,i}\right) \right]. \tag{25}$$

The linear term represents the local response around an operating point, while the saturating term describes bounded nonlinear effects without introducing an unrestricted neural structural equation. The tiered candidate graph ensures that all admissible edges follow the predefined physical ordering and therefore remain acyclic.

Let $o_{k,n,j} \in \{0,1\}$ indicate whether node j is observed in the current window, and let $\chi_j \in \{0,1\}$ indicate whether node j has at least one admissible parent and is therefore reconstructable. The effective structural weight is $\omega_{k,n,j,r} = \pi_{k,n,r} o_{k,n,j} \chi_j$. For a local minibatch $\mathcal{B}$, the mechanism-weighted structural loss is

$$\mathcal{L}_k^{\mathrm{sem}} = \frac{\sum_{n \in \mathcal{B}} \sum_{r=1}^{R} \sum_{j=1}^{J} \omega_{k,n,j,r} \left(\widehat{s}_{k,n,j}^{(r)} - s_{k,n,j} \right)^2}{\sum_{n \in \mathcal{B}} \sum_{r=1}^{R} \sum_{j=1}^{J} \omega_{k,n,j,r} + \varepsilon}, \tag{26}$$

where J is the number of retained physical nodes. Thus, a mechanism is trained primarily by samples with high membership support, while unobserved target nodes and nonreconstructable nodes do not contribute to its structural loss.

Before federated training, the edge probabilities and structural coefficients are initialized using sign-constrained robust regressions on the training data. The ridge regularization strength is selected by generalized cross-validation, while comparisons between reduced and complete structural equations provide BIC-based edge evidence. Let $g_{ij}^{0,(r)}$ denote the resulting edge-probability prior. Deviations from this prior are controlled by

$$\mathcal{L}_k^{\mathrm{edge}} = \frac{1}{R|\mathcal{E}_{\mathrm{phy}}|} \sum_{r=1}^{R} \sum_{(i,j) \in \mathcal{E}_{\mathrm{phy}}} \mathrm{KL}\left[\mathrm{Bern}\left(g_{ij}^{(r)}\right) \middle\| \mathrm{Bern}\left(g_{ij}^{0,(r)}\right) \right]. \tag{27}$$

This term discourages unsupported changes to the data-driven edge prior while retaining the ability to adapt edge probabilities when sufficient local evidence is available.

4) Causal Residual Fault Prediction: The physical state is interpretable but contains less diagnostic information than the complete temporal sensor window. SeaCausal-FL therefore does not replace the shared temporal classifier with the causal branch. Instead, all mechanisms share a nonlinear physical-state outcome network, while each mechanism learns only a zero-initialized linear residual.

Let $\mathbf{q}(\mathbf{s}_{k,n}) \in \mathbb{R}^C$ denote the logits produced by the shared causal outcome network. The causal logits associated with mechanism r are

$$\mathbf{z}_{k,n}^{(r)} = \mathbf{q}\left(\mathbf{s}_{k,n}\right) + \mathbf{W}_r^{\mathrm{res}} \mathbf{s}_{k,n} + \mathbf{b}_r^{\mathrm{res}}. \tag{28}$$

The residual parameters $\mathbf{W}_r^{\mathrm{res}}$ and $\mathbf{b}_r^{\mathrm{res}}$ are initialized to zero. Mechanism specialization is therefore introduced gradually during local optimization rather than imposed at initialization. The mechanism-specific outputs are combined using (5) and added to the shared diagnostic logits through the zero-initialized bounded scale in (6).

Since the causal contribution to the fused logits is initially zero, the causal branch receives an independent auxiliary classification signal:

$$\mathcal{L}_k^{\mathrm{causal}} = \frac{1}{|\mathcal{B}|} \sum_{n \in \mathcal{B}} \ell_{\mathrm{cw,ls}}\left(\mathbf{z}_{k,n}^{\mathrm{causal}}, y_{k,n}\right), \tag{29}$$

where $\ell_{\mathrm{cw,ls}}(\cdot)$ denotes the class-weighted cross-entropy computed from logits with the same label-smoothing setting as the primary classification loss.

Fault classification remains the primary objective under class imbalance. Let $\mathcal{A} = \{\mathrm{causal}, \mathrm{sem}, \mathrm{edge}, \mathrm{context}, \mathrm{drift}, \mathrm{mix}\}$ denote the active auxiliary-loss set, corresponding to causal classification, structural reconstruction, edge-prior regularization, context negative log-likelihood, fuzzy-prior drift, and mixture-prior KL divergence, respectively. Their relative contributions are automatically learned through log-variance parameters:

$$\mathcal{L}_k = \mathcal{L}_k^{\mathrm{cls}} + \frac{1}{|\mathcal{A}|} \sum_{m \in \mathcal{A}} \left[\frac{1}{2} \exp(-v_m) \mathcal{L}_{k,m} + \frac{1}{2} v_m \right], \tag{30}$$

where v_m is a trainable log-variance parameter constrained to a finite numerical range during optimization. The primary loss $\mathcal{L}_k^{\mathrm{cls}}$ is computed from the fused logits $\mathbf{z}_{k,n}$ and remains outside the automatic loss balancer. Its coefficient is therefore fixed at one and cannot be reduced by the uncertainty-based balancing mechanism. This design avoids manually assigning a separate coefficient to every fuzzy and causal auxiliary objective.

D. Federated Mechanism Coordination

After local training, mechanisms with the same index may no longer represent the same operating condition at different clients. Directly averaging such parameters can mix fuzzy prototypes and causal structures with different physical meanings. Before uploading its locally updated model, each selected client therefore aligns its mechanisms with the broadcast global reference. The aligned parameters and mechanism-related evidence statistics are subsequently used for server

aggregation. The coordination procedure consists of context–structure prealignment, conditional intervention–response refinement, and parameter-group-specific aggregation.

1) Context–Structure Prealignment: The local model is initialized from the current global model at the beginning of each communication round. Most mechanisms therefore retain their original ordering after local optimization. Computing intervention responses for every client in every round would therefore introduce unnecessary computational overhead. SeaCausal-FL first performs a low-cost prealignment using the operating-context centers and causal structures already contained in the local and global models.

Let $\boldsymbol{\xi}_{k,r}^{t+1}$ denote the center of local mechanism r after client k completes its local update in round t, and let $\boldsymbol{\xi}_s^t$ denote the center of global mechanism s. The context-center distance is

$$D_{k,r,s}^{\mathrm{ctx}} = \left\| \boldsymbol{\xi}_{k,r}^{t+1} - \boldsymbol{\xi}_s^t \right\|_2 . \tag{31}$$

Let $\mathbf{B}_{k,r}^{t+1} \in \mathbb{R}^{J\times J}$ and $\mathbf{B}_s^t \in \mathbb{R}^{J\times J}$ denote the effective causal coefficient matrices of the local and global mechanisms, respectively. For an admissible edge $(i,j) \in \mathcal{E}_{\mathrm{phy}}$, the corresponding entry is defined as $B_{ij}^{(r)} = g_{ij}^{(r)} \left(a_{ij}^{(r)} + c_{ij}^{(r)} \right)$ and is set to zero otherwise. This matrix provides a compact summary of the edge probability and the linear and saturating structural coefficients. The structure distance is defined as

$$D_{k,r,s}^{\mathrm{str}} = \frac{1}{J} \left\| \mathbf{B}_{k,r}^{t+1} - \mathbf{B}_s^t \right\|_{\mathrm{F}} , \tag{32}$$

where J is the number of physical variables and $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm.

Since the two distance matrices may have different numerical scales, SeaCausal-FL applies positive-median scaling:

$$\mathrm{RScale}(\mathbf{D}) = \frac{\mathbf{D}}{\mathrm{median}\left\{ D_{r,s} : D_{r,s} > 0 \right\} + \varepsilon}. \tag{33}$$

When no positive entry exists, the scale is set to one. The low-cost prealignment matrix is then

$$\mathbf{C}_k^{(0)} = \frac{1}{2} \left[\mathrm{RScale}\left(\mathbf{D}_k^{\mathrm{ctx}} \right) + \mathrm{RScale}\left(\mathbf{D}_k^{\mathrm{str}} \right) \right]. \tag{34}$$

This operation removes the scale difference between the context and structure distances, while their contributions remain equally weighted.

Let $\mathfrak{S}_R$ denote the set of permutations of R mechanisms. The preliminary assignment is obtained from

$$\sigma_k^{(0)} = \arg\min_{\sigma\in\mathfrak{S}_R} \sum_{s=1}^{R} C_{k,\sigma(s),s}^{(0)}, \tag{35}$$

where $\sigma_k^{(0)}(s)$ is the local mechanism assigned to global mechanism s. The assignment is solved using the Hungarian algorithm. If $\sigma_k^{(0)}$ is the identity permutation, the local ordering is retained and the intervention-based refinement is skipped.

2) Intervention–Response Signature Matching: Similar fuzzy centers and causal coefficient matrices do not necessarily imply equivalent physical behavior. Small differences distributed across several causal paths can produce different downstream responses after intervention. SeaCausal-FL therefore evaluates intervention–response signatures only when the context–structure prealignment indicates a possible permutation.

The intervention grid $\mathcal{I}$ contains feasible values of actionable physical variables. These values are obtained from client-level empirical quantiles of the training windows and consolidated across clients without transmitting raw sensor observations. For each intervention $(a,v) \in \mathcal{I}$, physical variable a is assigned value v. The factual mechanism-specific exogenous residuals are retained, and the intervention effects are propagated through the downstream physical nodes.

Let $\{\mathcal{B}_{k,b}^{\mathrm{cal}}\}_{b=1}^{B_k}$ denote the private calibration minibatches used for mechanism matching, where B_k is bounded by the configured signature-batch budget. For $u \in \{\mathrm{loc}, \mathrm{ref}\}$, representing the locally updated model and the broadcast global reference model, respectively, define the robust response of mechanism r under intervention (a,v) as

$$\boldsymbol{\delta}_{k,r}^{u}(a,v) = \frac{1}{B_k} \sum_{b=1}^{B_k} \mathrm{median}_{n\in\mathcal{B}_{k,b}^{\mathrm{cal}}} \left[\mathbf{s}_{k,n}^{\mathrm{cf},u,r}(a \leftarrow v) - \mathbf{s}_{k,n} \right]. \tag{36}$$

The intervention–response signature is then obtained by concatenating the responses over all interventions:

$$\boldsymbol{\psi}_{k,r}^{u} = \mathrm{vec}_{(a,v)\in\mathcal{I}} \left[\boldsymbol{\delta}_{k,r}^{u}(a,v) \right]. \tag{37}$$

Thus, each signature concatenates the minibatch-averaged median physical-state changes produced by all interventions. The locally updated model and the global reference are evaluated on the same private calibration minibatches, ensuring that their response signatures are comparable. The signature does not use fault labels and is therefore valid even when a client contains only a subset of the fault classes.

The intervention–response distance between local mechanism r and global mechanism s is

$$D_{k,r,s}^{\mathrm{resp}} = \left\| \boldsymbol{\psi}_{k,r}^{\mathrm{loc}} - \boldsymbol{\psi}_{k,s}^{\mathrm{ref}} \right\|_2 . \tag{38}$$

For the complete SeaCausal-FL model, the final alignment cost is

$$\mathbf{C}_k = \frac{1}{3} \left[\mathrm{RScale}\left(\mathbf{D}_k^{\mathrm{ctx}} \right) + \mathrm{RScale}\left(\mathbf{D}_k^{\mathrm{str}} \right) + \mathrm{RScale}\left(\mathbf{D}_k^{\mathrm{resp}} \right) \right]. \tag{39}$$

The final assignment is obtained as

$$\sigma_k^t = \arg\min_{\sigma\in\mathfrak{S}_R} \sum_{s=1}^{R} C_{k,\sigma(s),s}. \tag{40}$$

The fuzzy parameters, mechanism-indexed fuzzy priors, causal parameters, mechanism-specific output parameters, and mechanism-related evidence statistics are reordered according to σ_k^t. Shared parameters and class-count statistics are not permuted. Only the aligned model parameters and compact evidence summaries are used for aggregation; raw calibration windows and sample-level fault labels remain at the client.

3) Server-side Evidence-Aware Aggregation: Conventional sample-count averaging assumes that every local sample provides equal evidence for every model parameter. This assumption is unsuitable for SeaCausal-FL because a client may strongly activate only a subset of operating mechanisms and may contain no samples from some fault classes. The

server therefore selects the aggregation weight according to the parameter group being updated.

During local training, the midpoint fuzzy weights and fault labels already available in each forward pass are used to construct three complementary evidence statistics without an additional traversal of the local dataset. Let $\mathcal{T}_k^t$ denote the multiset of local sample occurrences processed by client k during communication round t, including their repeated occurrences across local epochs. For mechanism r and fault class c, the mechanism, mechanism–class, and class evidence are defined as

$$\begin{aligned} m_{k,r}^t &= \sum_{n\in\mathcal{T}_k^t} \pi_{k,n,r}, \\ m_{k,r,c}^t &= \sum_{n\in\mathcal{T}_k^t} \pi_{k,n,r}\mathbb{I}\left\{y_{k,n}=c\right\}, \\ n_{k,c}^t &= \sum_{n\in\mathcal{T}_k^t} \mathbb{I}\left\{y_{k,n}=c\right\}. \end{aligned} \tag{41}$$

Here, $m_{k,r}^t$ measures the effective support for mechanism r, $m_{k,r,c}^t$ measures the class-c evidence observed under that mechanism, and $n_{k,c}^t$ records the overall evidence for class c. The mechanism-indexed statistics $m_{k,r}^t$ and $m_{k,r,c}^t$ are reordered together with the corresponding mechanism parameters after alignment, whereas $n_{k,c}^t$ is unaffected by the permutation.

Let $\widetilde{\boldsymbol{\vartheta}}_k^{t+1}$ denote an aligned local parameter block and let w_k denote its corresponding evidence weight. The server aggregation operator is

$$\begin{aligned} &\mathrm{Agg}\left(\left\{\widetilde{\boldsymbol{\vartheta}}_k^{t+1}, w_k\right\}_{k\in\mathcal{S}_t}; \boldsymbol{\vartheta}^t\right) \\ &= \begin{cases} \displaystyle\sum_{k\in\mathcal{S}_t} \frac{w_k}{\sum_{j\in\mathcal{S}_t} w_j}\widetilde{\boldsymbol{\vartheta}}_k^{t+1}, & \displaystyle\sum_{j\in\mathcal{S}_t} w_j > 0, \\ \boldsymbol{\vartheta}^t, & \displaystyle\sum_{j\in\mathcal{S}_t} w_j = 0. \end{cases} \end{aligned} \tag{42}$$

Thus, when no participating client provides evidence for a parameter block, the corresponding parameter from the previous global model is retained.

Shared non-class parameters, including the temporal encoder, shared hidden layers, and other common parameters, use $w_k = N_k$. The class-c rows of the diagnostic output layer and the shared causal output classifier use $w_k = n_{k,c}^t$. The trainable parameters associated with fuzzy mechanism r and its structural causal model use $w_k = m_{k,r}^t$. Finally, the class-c output parameters specific to mechanism r use $w_k = m_{k,r,c}^t$.

The resulting global model exploits the evidence available from all participating clients without allowing a client with weak mechanism activation or a missing fault class to influence an unsupported parameter group. This coordination preserves the correspondence and evidence support of the fuzzy causal mechanisms while retaining sample-count-based sharing for the common diagnostic representation.

E. Interval Counterfactual Fault Reasoning

After federated training, the learned structural equations allow SeaCausal-FL to estimate how the predicted fault risk changes under a feasible physical intervention. The causal interpretation assumes that the physics-constrained candidate graph provides an approximately valid causal ordering for the retained physical variables, that unmeasured confounding does not dominate the modeled relations, and that the structural mechanisms remain stable within a counterfactual query. Accordingly, causal effects estimated from the real observational data are interpreted as model-based intervention analyses rather than observed causal ground truth; quantitative causal accuracy is evaluated on the real-data-calibrated semi-synthetic benchmark, where the underlying structure and intervention outcomes are known. The counterfactual procedure follows the abduction–action–prediction paradigm while retaining the factual temporal representation and operating-mechanism assignment.

1) Abduction of Exogenous Residuals: A counterfactual prediction should describe the same factual sample under a different action. Let $\mathcal{F}_j^{(r)}(\mathbf{s})$ denote the deterministic structural function of node j under mechanism r, corresponding to the right-hand side of (25) without the exogenous residual. For each mechanism, SeaCausal-FL infers the sample-specific residual as $\mathbf{u}_{k,n}^{(r)} = \mathbf{s}_{k,n} - \widehat{\mathbf{s}}_{k,n}^{(r)}$. Equivalently, for an endogenous node j,

$$u_{k,n,j}^{(r)} = s_{k,n,j} - \mathcal{F}_j^{(r)}\left(\mathbf{s}_{k,n}\right). \tag{43}$$

The residual $\mathbf{u}_{k,n}^{(r)}$ represents the sample-specific variation that is not explained by the deterministic structural equations. It is retained during counterfactual propagation so that the factual and counterfactual states correspond to the same underlying sample.

2) Action and Causal Propagation: Consider the intervention $\mathrm{do}(s_{k,n,a} = v)$, where a belongs to the actionable set comprising engine speed, water-brake load, charge-air intercooler cooling-water flow, and engine cooling-water flow. These variables correspond to controllable operating or cooling conditions in the considered system. Their feasible intervention values are obtained from the training-only empirical quantiles $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ rather than manually specified ranges. The mechanism assignment is anchored to the factual operating context throughout a counterfactual query. Therefore, interventions on context-defining variables are interpreted as within-mechanism perturbations rather than transitions to a different operating mechanism.

$$s_{k,n,j}^{\mathrm{cf},(r)} = \begin{cases} v, & j = a, \\ s_{k,n,j}, & j \notin \mathrm{De}(a)\cup\{a\}, \\ \mathcal{F}_j^{(r)}\left(\mathbf{s}_{k,n}^{\mathrm{cf},(r)}\right) + u_{k,n,j}^{(r)}, & j \in \mathrm{De}(a). \end{cases} \tag{44}$$

Variables that are neither intervened upon nor causally downstream of the intervention retain their factual values. Equation (44) produces one counterfactual physical state $\mathbf{s}_{k,n}^{\mathrm{cf},(r)}$ for each mechanism.

The mechanism-specific counterfactual causal logits are obtained by replacing the factual state in (28) with $\mathbf{s}_{k,n}^{\mathrm{cf},(r)}$, giving

$$\mathbf{z}_{k,n}^{\mathrm{cf},(r)} = \mathbf{q}\left(\mathbf{s}_{k,n}^{\mathrm{cf},(r)}\right) + \mathbf{W}_r^{\mathrm{res}}\mathbf{s}_{k,n}^{\mathrm{cf},(r)} + \mathbf{b}_r^{\mathrm{res}}. \tag{45}$$

The shared temporal logits $\mathbf{z}_{k,n}^{\mathrm{base}}$ remain factual because the intervention is applied only to the physical causal state rather than to a synthetically generated sensor window.

3) Interval Risk Estimation: The mechanism assignment is determined from the factual operating context and is not recomputed after intervention. This prevents an intervention from artificially reassigning the sample to another operating mechanism.

Let $\underline{\boldsymbol{\pi}}_{k,n}$, $\boldsymbol{\pi}_{k,n}$, and $\overline{\boldsymbol{\pi}}_{k,n}$ denote the factual lower, midpoint, and upper mechanism-weight vectors. The feasible IT2 weight set is

$$\mathcal{W}_{k,n} = \left\{ \mathbf{w} \in \mathbb{R}_+^R \;\middle|\; \begin{array}{l} \mathbf{1}^{\mathsf{T}}\mathbf{w} = 1, \\ \underline{\boldsymbol{\pi}}_{k,n} \preceq \mathbf{w} \preceq \overline{\boldsymbol{\pi}}_{k,n} \end{array} \right\}. \tag{46}$$

where $\preceq$ denotes componentwise inequality.

For each mechanism, the counterfactual causal logits are first fused with the factual temporal logits and mapped to a class-probability vector. Specifically,

$$\mathbf{p}_{k,n}^{\mathrm{cf},(r)} = \mathcal{S}\left(\mathbf{z}_{k,n}^{\mathrm{base}} + \lambda_{\mathrm{c}}\mathbf{z}_{k,n}^{\mathrm{cf},(r)}\right), \tag{47}$$

where $\mathcal{S}(\cdot)$ denotes the softmax mapping. For a feasible mechanism-weight vector $\mathbf{w} \in \mathcal{W}_{k,n}$, the counterfactual probability vector is

$$\mathbf{p}_{k,n}^{\mathrm{cf}}(\mathbf{w}) = \sum_{r=1}^{R} w_r \mathbf{p}_{k,n}^{\mathrm{cf},(r)}. \tag{48}$$

Let $p_{k,n,c}^{\mathrm{cf}}(\mathbf{w})$ denote the cth element of $\mathbf{p}_{k,n}^{\mathrm{cf}}(\mathbf{w})$. The nominal and interval-valued counterfactual risks are

$$\begin{aligned} p_{k,n,c}^{\mathrm{cf,mid}} &= p_{k,n,c}^{\mathrm{cf}}\left(\boldsymbol{\pi}_{k,n}\right), \\ \underline{p}_{k,n,c}^{\mathrm{cf}} &= \min_{\mathbf{w}\in\mathcal{W}_{k,n}} p_{k,n,c}^{\mathrm{cf}}(\mathbf{w}), \\ \overline{p}_{k,n,c}^{\mathrm{cf}} &= \max_{\mathbf{w}\in\mathcal{W}_{k,n}} p_{k,n,c}^{\mathrm{cf}}(\mathbf{w}). \end{aligned} \tag{49}$$

Here, $p_{k,n,c}^{\mathrm{cf,mid}}$ is the nominal counterfactual risk, while $[\underline{p}_{k,n,c}^{\mathrm{cf}}, \overline{p}_{k,n,c}^{\mathrm{cf}}]$ characterizes the variation induced by uncertainty in the factual operating-mechanism assignment. It is therefore interpreted as a mechanism-uncertainty interval rather than a nominal-coverage prediction interval unless an additional calibration procedure is applied.

IV. Experiments

A. Experimental Setup

1) Datasets and Federated Protocol: We evaluate SeaCausal-FL on the Marine Engine Fault Dataset [43]. Its 15 fault–load recordings are treated as 15 virtual maritime IoT clients and form a six-class task containing normal operation and five fault types. Each recording is chronologically divided into 60%, 20%, and 20% training, validation, and test segments, with a 32-sample guard interval between adjacent segments. Windows of length $T = 32$ are extracted with a stride of 8, yielding 6,494/2,018/2,129 windows under the natural partition. Missing values are mean-imputed and accompanied by a binary mask.

Besides the natural partition, we construct an IID partition and Dirichlet label-skew partitions with $\alpha \in \{0.5, 1.0\}$. For causal and counterfactual evaluation, we further construct a real-data-calibrated semi-synthetic benchmark from the training portion of the marine-engine dataset. The benchmark retains the 15-client structure and 18 selected physical variables, while the empirical operating ranges, marginal statistics, and fault-related response patterns of the real training data are used to calibrate the synthetic generation process. Mechanism-specific structural causal models generate the physical states and provide known causal coefficient matrices, which are retained as ground truth for causal-structure evaluation. This benchmark is formulated as a binary fault-risk task, and the generated samples are divided into training, validation, and test subsets at the client level. For each factual sample, interventions are applied to predefined actionable variables over feasible values derived from the calibrated operating ranges, and the same generating structural causal model (SCM) is used to obtain the corresponding counterfactual states and fault risks. The generated data are additionally checked against the real-data calibration statistics before causal evaluation. The real dataset is therefore used for the main six-class diagnostic experiments, whereas the real-data-calibrated semi-synthetic benchmark is used only when causal or counterfactual ground truth is required.

2) Compared Methods: For diagnostic evaluation, SeaCausal-FL is compared with six representative federated baselines: a federated TSK fuzzy baseline (FedFuzzy-TSK) implemented following the federated TSK learning setting in [44], FedProx [45], FedAdam [46], Ditto [47], FedBN [48], and MOON [49]. These methods represent fuzzy federated learning, proximal optimization, adaptive server optimization, personalized federated learning, local normalization for feature-shifted non-IID data, and representation-level contrastive learning, respectively. The same temporal backbone and data partitions are used whenever applicable to provide a consistent comparison.

For the causal and counterfactual evaluation on the semi-synthetic benchmark, SeaCausal-FL is further compared with Pooled-SEM, Local-SEM, and Oracle-SEM, which are constructed following standard structural causal modeling principles [50]. Pooled-SEM learns a single structural equation model from pooled training samples, Local-SEM estimates separate causal models from individual clients, and Oracle-SEM fits mechanism-specific structural models using the ground-truth operating-mechanism assignments. These baselines respectively represent centralized causal estimation, fully local causal learning, and causal modeling with oracle regime information.

3) Implementation Details: The common backbone concatenates the 69 sensor channels with their missing-value masks, projects the resulting input to 96 channels, and applies two residual one-dimensional convolutional layers with a kernel size of 3, eight-group normalization, GELU activation, and dropout. The average and last-step temporal summaries are combined into a 96-dimensional representation. The diag-

nostic head contains a 96-unit hidden layer and produces six class logits. FedBN replaces group normalization with client-specific batch normalization. SeaCausal-FL additionally uses three operating-context variables, seven automatically selected fuzzy mechanisms, and an 18-node SCM. Its causal outcome branch contains a shared hidden layer with 96 units and six output logits, together with a mechanism-specific linear residual that maps the 18 physical variables to the six fault classes.

Federated training uses at most 80 rounds, 60% client participation, two local epochs, and a batch size of 128. AdamW is used with learning rate 8×10^{-4}, weight decay 10^{-4}, gradient clipping at 5, and dropout 0.2. The checkpoint with the highest validation F1-score is used for testing. All fuzzy and causal priors and the intervention quantiles $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ are estimated from training data only.

4) Evaluation Metrics: Macro-F1 score is used as the primary diagnostic metric, together with accuracy, precision, recall, AUROC, and AUPRC to evaluate classification and ranking performance. The leave-one-load-out experiment further reports accuracy, precision, recall, and Macro-F1 score to assess generalization to unseen operating conditions. The leave-one-fault-type-out experiment reports the same metrics on the remaining known fault classes after one fault type is completely excluded from training, thereby evaluating diagnostic robustness to incomplete fault-type coverage in the training data. On the real-data-calibrated semi-synthetic causal benchmark, causal structure recovery is evaluated using Edge-F1, Edge-AUPRC, and coefficient RMSE. Counterfactual fidelity is assessed by counterfactual risk MAE, PEHE, and effect-sign accuracy, while policy regret and action accuracy are used to evaluate intervention quality.

B. Performance Evaluation of SeaCausal-FL Framework

Table I reports the diagnostic results under four client partitions with different degrees of data heterogeneity. SeaCausal-FL achieves the highest F1-score under IID, natural non-IID, and Dirichlet $\alpha = 1.0$, reaching 84.42%, 90.48%, and 85.61%, respectively. Under Dirichlet $\alpha = 0.5$, FedAdam obtains the highest F1-score of 88.82%, while SeaCausal-FL achieves 87.76%, with a difference of only 1.06 percentage points. When the four partitions are considered together, SeaCausal-FL reaches average accuracy, precision, recall, and F1 values of 86.51%, 86.84%, 91.23%, and 87.07%, respectively. Compared with the strongest baseline for each corresponding metric, the improvements are 2.33, 3.51, 4.24, and 3.67 percentage points. Similar gains are observed for AUROC and AUPRC, where SeaCausal-FL achieves 98.98% and 94.81%. In particular, the improvement in recall indicates that fewer fault samples are missed, which is important for marine-engine monitoring where an undetected abnormal condition can lead to a more serious operational consequence.

The convergence behavior in Fig. 3 provides a further view of the effect of client heterogeneity. Under IID, most methods reach a relatively stable region after the early communication rounds, and their performance differences are moderate. The gap becomes more evident under the natural non-IID setting,

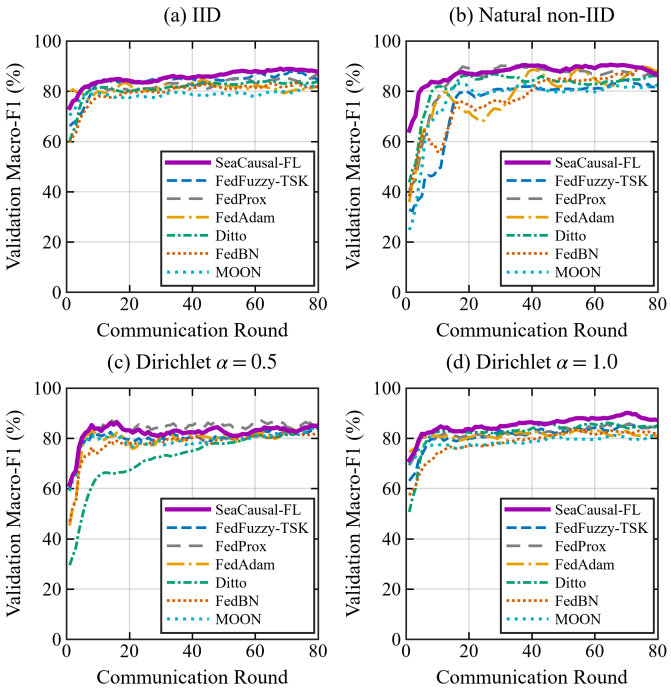

Fig. 3. Comparison of convergence performance across different models under four client partitions.

where several baselines exhibit larger fluctuations during the first part of training. SeaCausal-FL rises rapidly and remains in a high F1 region after convergence. Similar behavior is observed under the two Dirichlet partitions. In the $\alpha = 0.5$ case, several methods remain competitive and FedAdam eventually obtains a slightly higher F1-score, which is consistent with the results in Table I. However, SeaCausal-FL maintains a relatively stable trajectory as the partition changes. This is particularly relevant to federated marine-engine diagnosis because the composition of participating clients and their local class distributions can vary substantially across communication rounds.

The leave-one-load-out experiment further examines whether the learned model remains effective when the operating condition itself is unseen during training. For each setting in Fig. 4, all samples belonging to one load level are removed from the training set and used only for evaluation. This setting differs from ordinary client heterogeneity because the model must transfer its diagnostic knowledge to an operating region that has not contributed to model optimization. SeaCausal-FL maintains strong accuracy, precision, recall, and F1-score across the four held-out loads, whereas the relative performance of the competing methods changes more noticeably with the excluded load. The difference becomes particularly clear when the 85% load is held out, where several baselines show a substantial decrease across multiple metrics while SeaCausal-FL retains a comparatively high diagnostic performance.

The advantage under unseen loads is closely related to the treatment of operating conditions in the proposed framework. A diagnostic model trained only from statistical associations can associate a fault with sensor patterns that are specific to

TABLE I
DIAGNOSTIC PERFORMANCE COMPARISON OF DIFFERENT MODELS UNDER FOUR CLIENT PARTITIONS.

Method	Partition-wise F1-score↑				Average over four partitions					
	IID	Natural Non-IID	Dir. ($\alpha = 0.5$) Non-IID	Dir. ($\alpha = 1.0$) Non-IID	Acc.↑	Prec.↑	Rec.↑	F1-score↑	AUROC↑	AUPRC↑
FedFuzzy-TSK	82.66	88.95	79.40	79.73	81.72	83.08	86.99	82.69	97.33	90.74
FedProx	77.36	89.94	86.82	77.74	83.03	83.33	86.75	83.22	98.23	91.34
FedAdam	80.44	87.03	**88.82**	77.30	84.18	83.11	85.27	83.40	97.74	88.57
Ditto	81.70	86.65	76.12	80.94	80.77	81.41	86.53	81.35	97.31	88.29
FedBN	79.25	85.16	79.59	79.67	80.01	80.71	84.61	80.92	97.11	87.39
MOON	75.15	83.31	86.89	76.00	79.61	81.10	84.32	80.34	96.38	86.91
SeaCausal-FL	**84.42**	**90.48**	87.76	**85.61**	**86.51**	**86.84**	**91.23**	**87.07**	**98.98**	**94.81**

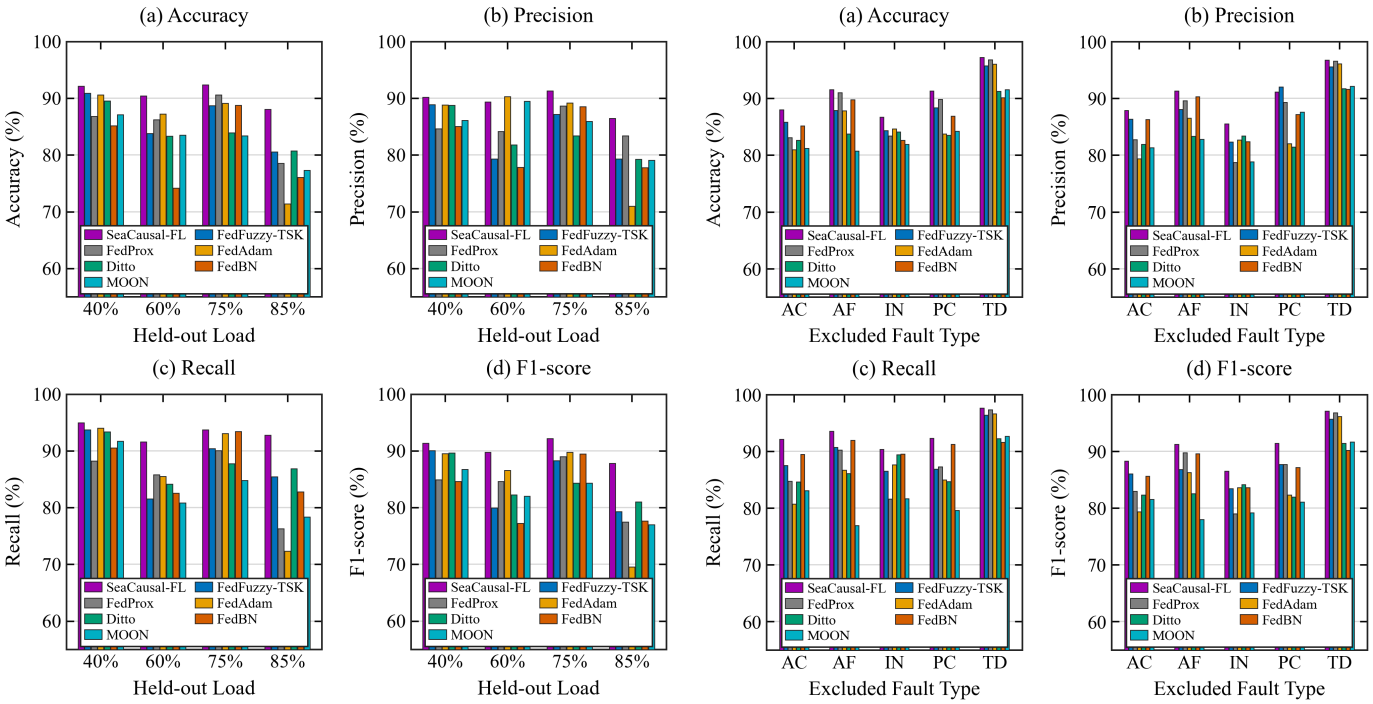

Fig. 4. Diagnostic performance under leave-one-load-out training.

Fig. 5. Diagnostic performance under leave-one-fault-type-out training.

the loads represented in the training clients. Once the operating point changes, the same fault can produce a different physical response and the learned decision boundary becomes less reliable. SeaCausal-FL instead represents the operating context through overlapping fuzzy mechanisms and associates each mechanism with its own causal response model. A sample located between two operating regions can therefore receive support from both mechanisms rather than being assigned to a fixed load interval. This soft operating representation reduces abrupt changes between neighboring conditions and provides a smoother basis for transferring the learned physical relations to an unseen load.

A different form of training-data heterogeneity is examined through the leave-one-fault-type-out evaluation in Fig. 5. In each setting, all samples belonging to one fault type are excluded from model training, and diagnostic performance is evaluated on the remaining known fault classes. This experiment therefore examines the sensitivity of federated diagnosis to incomplete fault-type coverage rather than recognition of the excluded fault itself. The performance of several baselines varies noticeably as different fault types are removed, indicating that the composition of fault categories available during training can substantially affect the learned representation.

SeaCausal-FL remains consistently competitive across all five fault-omission settings and shows smaller performance degradation in the more challenging cases. The advantage is observed across accuracy, precision, recall, and F1-score rather than being confined to a single metric, suggesting that the proposed representation is less sensitive to the absence of a particular fault category during training.

This behavior can be attributed to the complementary roles of the two diagnostic paths. The shared temporal encoder extracts fault-discriminative information that can be reused across clients, while the fuzzy causal branch describes how changes in operating state are related to downstream physical responses. The final prediction therefore does not rely solely on a fault pattern observed under a particular client or load condition. Such a representation is less sensitive to changes in the fault-category composition of the training data and helps preserve diagnostic performance when some fault types are unavailable during model optimization.

TABLE II
ABLATION RESULTS UNDER THE NATURAL NON-IID PARTITION.

Variant	Acc.↑	Prec.↑	Rec.↑	F1-score↑	AUROC↑	AUPRC↑
w/o IT2	83.33	82.90	87.47	82.62	98.11	87.50
Single mechanism ($R = 1$)	88.40	86.98	90.13	88.03	98.82	91.64
w/o SEM	84.78	84.33	88.95	84.36	98.19	87.70
w/o edge-prior regularization	84.78	84.41	89.17	84.39	98.28	87.91
w/o causal residual	84.08	83.51	88.24	83.48	98.16	87.60
w/o alignment	86.38	85.69	89.20	86.44	98.04	91.68
SeaCausal-FL	**90.28**	**89.84**	**93.39**	**90.48**	**99.02**	**92.48**

C. Ablation and Sensitivity Analyses

Table II evaluates the contribution of the main components under the natural non-IID partition. Removing the IT2 fuzzy modeling reduces F1-score from 90.48% to 82.62% and AUPRC from 92.48% to 87.50%. This confirms that the fuzzy layer is not used only for uncertainty estimation; it also represents overlapping operating conditions through soft mechanism memberships rather than fixed regime boundaries [26, 27]. Removing the SEM loss or edge-prior regularization produces similar reductions, with F1-score decreasing to 84.36% and 84.39%, respectively. The former weakens the structural modeling of physical relations, while the latter weakens the constraint that keeps learned edge probabilities close to their data-driven structural priors. Their comparable degradation indicates that both structural reconstruction and selective causal relations are important for the causal branch [28, 29].

The context-conditioned causal residual also has a pronounced effect. Without this component, F1-score decreases to 83.48%, indicating that operating-dependent causal information needs to directly refine the shared diagnostic output. In comparison, using a single causal mechanism gives a smaller reduction to 88.03%. This result suggests that one common mechanism already captures part of the shared physical structure, but multiple mechanisms remain useful for describing operating-dependent responses. Without mechanism alignment, F1-score decreases to 86.44%. Local mechanisms can drift toward different operating regions during training, and direct aggregation can consequently combine components with different physical meanings. The result supports mechanism-level coordination before federated aggregation, consistent with the need to coordinate causal knowledge learned from distributed data [39, 40].

It is also noteworthy that AUROC remains above 98% for all ablation variants, whereas F1-score and AUPRC change more substantially. This indicates that the shared temporal path preserves strong overall class separability, while the fuzzy causal components mainly improve the reliability of the final decisions under heterogeneous operating conditions.

The effect of client availability is examined in Fig. 6(a). F1-score remains approximately between 85% and 90% as the participation ratio varies from 20% to 100%, with the highest value observed around 60%. The nonmonotonic trend reflects the tradeoff between information coverage and heterogeneous local updates: fewer participating clients increase round-to-round variation, whereas more clients introduce a broader range of local distributions. More importantly, the overall variation remains moderate, showing that SeaCausal-FL does not rely on full client participation in every communication round.

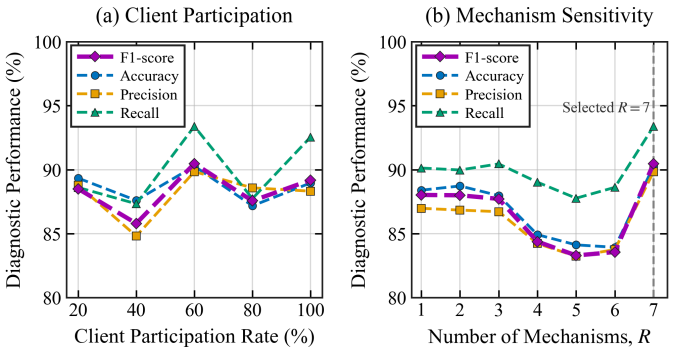

Fig. 6. Performance under different client participation ratios and causal-mechanism numbers.

The number of fuzzy causal mechanisms shows a stronger influence in Fig. 6(b). With $R = 1$–3, F1-score remains close to 88%, while $R = 4$–6 produces a noticeable decrease. Dividing the operating space into more mechanisms can reduce the effective samples available for mechanism-specific structural estimation when the resulting decomposition is not sufficiently informative. The sensitivity analysis evaluates the complete set of mechanism numbers satisfying the sample-sufficiency criterion rather than manually truncating the search at $R = 7$. For the present training data, $R = 1, \ldots, 7$ remain eligible, while larger values are excluded because at least one fitted component does not provide sufficient soft or hard sample support for reliable structural estimation. Within this admissible set, $R = 7$ is selected by the data-driven criterion and also provides the strongest diagnostic performance. This agreement supports determining the mechanism number from the available operating evidence rather than prescribing a fixed number of regimes.

D. Causal Interpretation and Counterfactual Reasoning

The causal evaluation is conducted on the real-data-calibrated semi-synthetic benchmark, where the ground-truth causal structure and intervention outcomes are available for evaluation. This setting allows diagnostic performance, causal structure recovery, counterfactual estimation, and intervention quality to be examined within the same controlled environment. As shown in Fig. 8(a), SeaCausal-FL achieves the strongest factual diagnostic performance among the compared causal models in terms of accuracy, precision, recall, and F1-score. This result indicates that introducing mechanism-specific causal modeling does not compromise the shared diagnostic capability. Instead, the causal branch provides operating-dependent corrections while the shared temporal path preserves fault-discriminative information.

The learned causal structure is further examined qualitatively in Fig. 7 and quantitatively in Fig. 8(b). For the representative operating mechanism, the estimated coefficient matrix recovers the main sparse pattern and several dominant relations in the ground-truth matrix, although differences remain in

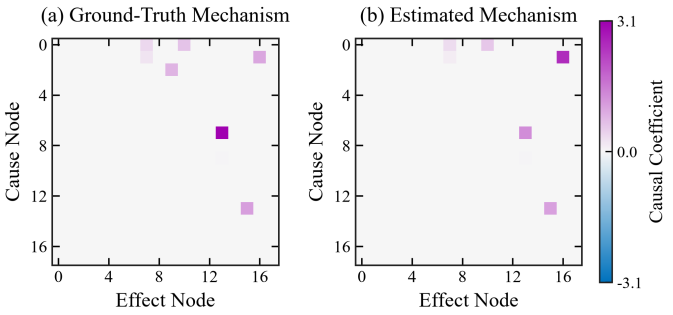

Fig. 7. Comparison between the ground-truth and estimated causal coefficient matrices for a representative mechanism.

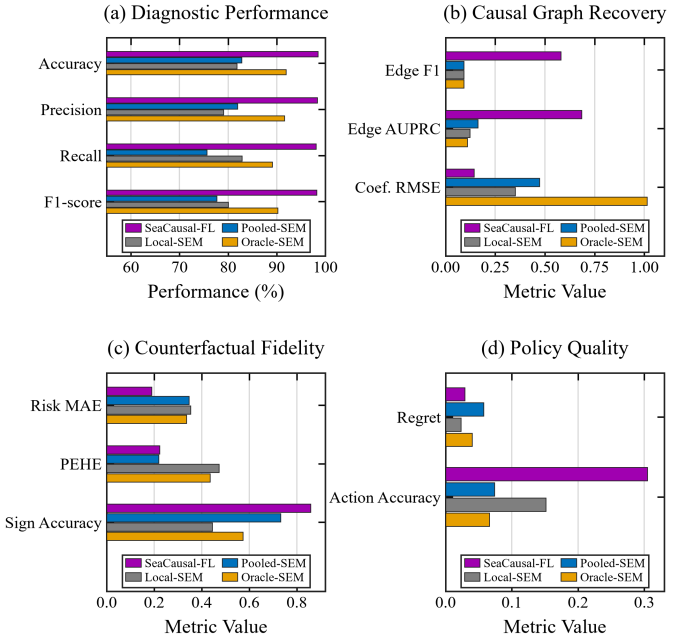

Fig. 8. Comparison of diagnostic performance, causal graph recovery, counterfactual fidelity, and policy quality across different methods.

individual coefficient magnitudes. SeaCausal-FL achieves an Edge-F1 of approximately 0.58 and an Edge-AUPRC of 0.68, while the compared SEM baselines remain below approximately 0.10 and 0.17, respectively. Its coefficient RMSE is also reduced to about 0.14. The higher edge-level accuracy and lower coefficient error indicate that the proposed model recovers a more selective and quantitatively accurate structural representation. This is important for subsequent intervention analysis because counterfactual propagation depends on both directed relations and their effect magnitudes [28, 29]. The mechanism-specific formulation also differs from federated causal approaches that mainly recover a single global causal structure from distributed data [39, 40].

The counterfactual results in Fig. 8(c) examine whether the recovered structural relations can reproduce intervention effects at the sample level. SeaCausal-FL obtains the lowest counterfactual risk MAE among the compared methods and remains competitive in PEHE. It also achieves the highest effect-sign accuracy, indicating that the predicted direction of risk change agrees more frequently with the reference intervention effect. Correctly identifying whether an intervention increases or decreases fault risk is particularly relevant to maintenance reasoning. The larger PEHE of Local-SEM and Oracle-SEM further shows that fitting separate structural models to local data or known operating regimes alone does not necessarily provide accurate individual-level counterfactual effects.

The policy-level results in Fig. 8(d) further evaluate whether these counterfactual estimates support useful intervention selection. SeaCausal-FL achieves the highest action accuracy, indicating that its selected intervention agrees with the reference optimal action more frequently than those of the SEM baselines. Its policy regret is also lower than those of Pooled-SEM and Oracle-SEM and remains close to the lowest value obtained by Local-SEM. The difference between these two metrics is expected: action accuracy evaluates whether the exact optimal action is selected, whereas policy regret measures the excess cost associated with the selected action. A method can therefore obtain relatively low regret by selecting a near-optimal action even when it does not exactly match the reference optimum. Overall, SeaCausal-FL maintains a favorable balance between causal structure recovery, sample-level counterfactual estimation, and intervention selection.

V. Conclusion

This paper presented SeaCausal-FL for federated marine-engine fault diagnosis under operating-condition and label heterogeneity. The framework combines a shared temporal diagnostic model with data-driven interval type-2 fuzzy mechanisms and physics-constrained structural causal models. Mechanism alignment and evidence-aware aggregation preserve operating-specific knowledge across clients, while the learned structural equations enable intervention-based counterfactual reasoning. Experiments on the real marine-engine dataset showed strong performance across IID and non-IID partitions, unseen loads, and incomplete fault-type coverage. Ablation results verified the importance of fuzzy mechanism modeling, causal structural learning, residual correction, and mechanism alignment. On the real-data-calibrated causal benchmark, SeaCausal-FL also achieved more accurate causal-structure recovery and favorable counterfactual and policy-level performance. Future work will consider larger multi-vessel deployments and richer real intervention data. As a potential future direction, we are looking forward to extending our method to improve the performance of various applications, such as large language models [51–58] and split learning system [16, 59–63].

References

[1] S. Aslam, M. P. Michaelides, and H. Herodotou, "Internet of ships: A survey on architectures, emerging applications, and challenges," *IEEE Internet of Things Journal*, vol. 7, no. 10, pp. 9714–9727, 2020.

[2] H. Chen, Y. Wen, Y. Huang, C. Xiao, and Z. Sui, "Edge computing enabling internet of ships: A survey on architectures, emerging applications, and challenges," *IEEE Internet of Things Journal*, vol. 12, no. 2, pp. 1509–1528, 2025.

[3] Y. Huo, X. Dong, and S. J. Beatty, "Cellular communications in ocean waves for maritime internet of things," *IEEE Internet of Things Journal*, vol. 7, no. 10, pp. 9965–9979, 2020.

[4] A. Youssef, H. Noura, A. El Amrani, E. M. El Adel, and M. Ouladsine, "A survey on data-driven fault diagnostic techniques for marine diesel engines," *IFAC-PapersOnLine*, vol. 58, no. 4, pp. 55–60, 2024.

[5] J. Kowalski, B. Krawczyk, and M. Woźniak, "Fault diagnosis of marine 4-stroke diesel engines using a one-vs-one extreme learning ensemble," *Engineering Applications of Artificial Intelligence*, vol. 57, pp. 134–141, 2017.

[6] R. Wang, H. Chen, C. Guan, W. Gong, and Z. Zhang, "Research on the fault monitoring method of marine diesel engines based on the manifold learning and isolation forest," *Applied Ocean Research*, vol. 112, p. 102681, 2021.

[7] H. Yuan, Z. Chen, Z. Lin, J. Peng, Z. Fang, Y. Zhong, Z. Song, and Y. Gao, "SatSense: Multi-Satellite Collaborative Framework for Spectrum Sensing," *IEEE Trans. Cogn. Commun. Netw.*, 2025.

[8] J. Peng, Z. Chen, Z. Lin, H. Yuan, Z. Fang, L. Bao, Z. Song, Y. Li, J. Ren, and Y. Gao, "SUMS: Sniffing Unknown Multiband Signals under Low Sampling Rates," *IEEE Trans. Mobile Comput.*, 2024.

[9] Z. Lin, G. Zhu, Y. Deng, X. Chen, Y. Gao, K. Huang, and Y. Fang, "Efficient Parallel Split Learning over Resource-Constrained Wireless Edge Networks," *IEEE Trans. Mobile Comput.*, vol. 23, no. 10, pp. 9224–9239, 2024.

[10] X. Guan, Z. Zhang, Z. Lin, Z. Sun, T. Duan, Z. Fang, R. Wang, H. Cui, W. Ni, J. Luo, *et al.*, "Fluxshard: Motion-aware feature cache reuse for collaborative video analytics in mobile edge computing," *arXiv preprint arXiv:2605.06027*, 2026.

[11] Z. Lin, O. Aouedi, W. Ni, S. Chatzinotas, and X. Chen, "Gapsl: A gradient-aligned parallel split learning on heterogeneous data," *arXiv preprint arXiv:2603.18540*, 2026.

[12] Z. Shi, Z. Wang, Z. Yuan, M. Wang, Z. Liu, and J. Fei, "A universal transfer learning framework for cross-working-condition marine diesel engine fault diagnosis based on fine-tuning strategy," *Applied Energy*, vol. 392, p. 125962, 2025.

[13] Z. Zhao, Z. Jin, X. Xin, Y. Fu, X. Huang, L. Li, H. Qin, C. Wei, Y. Li, and Y. Liu, "Cross-domain fault diagnosis of marine diesel engines based on stepwise diffusion and iterative bidirectional optimization," *Engineering Applications of Artificial Intelligence*, vol. 155, p. 110994, 2025.

[14] Y. Gao, X. Zheng, J. Li, L. Zong, H. Yin, H. Li, and G. Lu, "Disentanglement learning with adaptive centroid alignment for multiple target domains fault diagnosis," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 9, pp. 10 779–10 790, 2024.

[15] Q. Qian, B. Zhang, C. Li, Y. Mao, and Y. Qin, "Federated transfer learning for machinery fault diagnosis: A comprehensive review of technique and application," *Mechanical Systems and Signal Processing*, vol. 223, p. 111837, 2025.

[16] Z. Lin, G. Qu, W. Wei, X. Chen, and K. K. Leung, "AdaptSFL: Adaptive split federated learning in resource-constrained edge networks," *IEEE Transactions on Networking*, vol. 33, no. 6, pp. 2993–3008, 2025.

[17] G. Chen, J. Li, Y. Cui, Q. Wu, Y. Ni, M. Hua, and S. Lu, "IRS aided federated learning: Multiple access and fundamental tradeoff," *IEEE Transactions on Mobile Computing*, vol. 25, no. 7, 2026.

[18] W. Zhang, X. Li, H. Ma, Z. Luo, and X. Li, "Federated learning for machinery fault diagnosis with dynamic validation and self-supervision," *Knowledge-Based Systems*, vol. 213, p. 106679, 2021.

[19] W. Zhang and X. Li, "Data privacy preserving federated transfer learning in machinery fault diagnostics using prior distributions," *Structural Health Monitoring*, vol. 21, no. 4, pp. 1329–1344, 2022.

[20] C. Zhao and W. Shen, "Federated domain generalization: A secure and robust framework for intelligent fault diagnosis," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 2, pp. 2662–2670, 2024.

[21] Y. Han, Z. Liu, Q. Huang, and Y. Zhang, "Class information-guided personalized federated learning for fault diagnosis under label distribution skew," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–12, 2024, art. no. 3541012.

[22] T. Cui, W. Dai, and H. Zhang, "Cooperative reinforced resilient federated learning for clients with multiple working conditions," *IEEE Transactions on Reliability*, vol. 75, pp. 1469–1482, 2026.

[23] D. Xu, M. Jia, T. Chen, Y. Liu, D. Chen, T. Chai, and T. Yang, "Federated open-set fault diagnosis for unknown bearing fault detection," *Mechanical Systems and Signal Processing*, vol. 246, p. 113917, 2026.

[24] Z. Lin, Z. Chen, X. Chen, W. Ni, and Y. Gao, "Hasfl: Heterogeneity-aware split federated learning over edge computing systems," *IEEE Transactions on Mobile Computing*, vol. 25, no. 8, pp. 12 455–12 471, 2026.

[25] Z. Lin, W. Wei, Z. Chen, C.-T. Lam, X. Chen, Y. Gao, and J. Luo, "Hierarchical split federated learning: Convergence analysis and system optimization," *IEEE Transactions on Mobile Computing*, vol. 24, no. 10, pp. 9352–9367, 2025.

[26] Q. Liang and J. M. Mendel, "Interval type-2 fuzzy logic systems: Theory and design," *IEEE Transactions on Fuzzy Systems*, vol. 8, no. 5, pp. 535–550, 2000.

[27] J. Qiao, Z. Sun, and X. Meng, "Interval type-2 fuzzy neural network based on active semi-supervised learning for non-stationary industrial processes," *IEEE Transactions on Automation Science and Engineering*, vol. 21, no. 2, pp. 1151–1162, 2024.

[28] J. Pearl, "Causal inference in statistics: An overview," *Statistics Surveys*, vol. 3, pp. 96–146, 2009.

[29] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio, "Toward causal representation learning," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 612–634, 2021.

[30] P. Han, A. L. Ellefsen, G. Li, F. T. Holmeset, and H. Zhang, "Fault detection with lstm-based variational autoencoder for maritime components," *IEEE Sensors Journal*, vol. 21, no. 19, pp. 21 903–21 912, 2021.

[31] R. F. Naryanto, M. K. Delimayanti, A. Naryaningsih, B. Warsuta, R. Adi, and B. A. Setiawan, "Diesel engine fault detection using deep learning based on lstm," in *2023 7th International Conference on Electrical, Telecommunication and Computer Engineering (ELTICOM)*. IEEE, 2023, pp. 37–42.

[32] R. Wang, H. Chen, and C. Guan, "DPGCN model: A novel fault diagnosis method for marine diesel engines based on imbalanced datasets," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–11, 2023.

[33] W. Zhang, R. Jiang, S. Chen, Z. Yuan, S. Liu, and J. Wang, "A dual-path cnn-mlp-cbam network for intelligent fault diagnosis of marine diesel engines," *IEEE Access*, vol. 14, pp. 31 881–31 893, 2026.

[34] X. Li, C. Zhang, X. Li, and W. Zhang, "Federated transfer learning in fault diagnosis under data privacy with target self-adaptation," *Journal of Manufacturing Systems*, vol. 68, pp. 523–535, 2023.

[35] J. Chen, J. Tang, and W. Li, "Industrial edge intelligence: Federated-meta learning framework for few-shot fault diagnosis," *IEEE Transactions on Network Science and Engineering*, vol. 10, no. 6, pp. 3561–3573, 2023.

[36] J. Cui, J. Li, Z. Mei, K. Wei, S. Wei, M. Ding, W. Chen, and S. Guo, "Federated meta-learning for few-shot fault diagnosis with representation encoding," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–12, 2023, art. no. 3536812.

[37] H. Fan, S. Liu, X. Cao, and X. Zhang, "A prototype-guided federated learning based fault diagnosis method of mechanical transmission system under label distribution skew," *Neurocomputing*, vol. 656, p. 131532, 2025.

[38] S. Lu, Z. Gao, P. Zhang, Q. Xu, T. Xie, and A. Zhang, "Event-triggered federated learning for fault diagnosis of offshore wind turbines with decentralized data," *IEEE Transactions on Automation Science and Engineering*, vol. 21, no. 2, pp. 1271–1283, 2024.

[39] D. Yang, X. He, J. Wang, G. Yu, C. Domeniconi, and J. Zhang, "Federated causality learning with explainable adaptive optimization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 15, 2024, pp. 16 308–16 315.

[40] X. Guo, K. Yu, L. Liu, and J. Li, "FedCSL: A scalable and accurate approach to federated causal structure learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 11, 2024, pp. 12 235–12 243.

[41] Y. Wang, F. Cao, K. Yu, and J. Liang, "Federated causal structure learning with non-identical variable sets," in *Proceedings of the 42nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 267. PMLR, 2025, pp. 62 445–62 466.

[42] N. Mohamed, A. Mohanty, and N. Gebraeel, "Federated causal representation learning in state-space systems for decentralized counterfactual reasoning," *arXiv preprint arXiv:2602.19414*, 2026.

[43] A. BahooToroody, O. Bondarenko, M. M. Abaei, N. Yoichi, and E. Zio, "Marine engine fault dataset: Open-access data under controlled reference and fault scenario conditions," *arXiv preprint arXiv:2607.19444*, 2026.

[44] J. L. Corcuera Bárcena, P. Ducange, F. Marcelloni, A. Renda, F. Ruffini, and A. Schiavo, "Federated TSK models for predicting quality of experience in B5G/6G networks," in *2023 IEEE International Conference on Fuzzy Systems (FUZZ)*. IEEE, 2023, pp. 1–8.

[45] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proceedings of Machine Learning and Systems*, vol. 2, 2020, pp. 429–450.

[46] S. J. Reddi, Z. Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konečný, S. Kumar, and H. B. McMahan, "Adaptive federated optimization," in *International Conference on Learning Representations*, 2021.

[47] T. Li, S. Hu, A. Beirami, and V. Smith, "Ditto: Fair and robust federated learning through personalization," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 139. PMLR, 2021, pp. 6357–6368.

[48] X. Li, M. Jiang, X. Zhang, M. Kamp, and Q. Dou, "FedBN: Federated

learning on non-iid features via local batch normalization," in *International Conference on Learning Representations*, 2021.

[49] Q. Li, B. He, and D. Song, "Model-contrastive federated learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 713–10 722.

[50] J. Peters, D. Janzing, and B. Schölkopf, *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 2017.

[51] Z. Fang, Z. Lin, Z. Chen, X. Chen, Y. Gao, and Y. Fang, "Automated Federated Pipeline for Parameter-Efficient Fine-Tuning of Large Language Models," *IEEE Trans. Mobile Comput.*, 2025.

[52] Y. Zhang, M. Hu, Z. Lin, X. Fan, F. Xie, Z. Fang, J. Yang, W. Zhu, Z. Chen, C. Lv, *et al.*, "Hera: Learning long-horizon coordination for device-cloud collaborative llm agents," *arXiv preprint arXiv:2605.24598*, 2026.

[53] Z. Lin, Y. Zhang, Z. Chen, Z. Fang, X. Chen, P. Vepakomma, W. Ni, J. Luo, and Y. Gao, "HSplitLoRA: A Heterogeneous Split Parameter-Efficient Fine-Tuning Framework for Large Language Models," *arXiv preprint arXiv:2505.02795*, 2025.

[54] Z. Fang, Q. Wang, H. An, Z. Lin, Y. Deng, X. Chen, and Y. Fang, "Aggregation alignment for federated learning with mixture-of-experts under data heterogeneity," *arXiv preprint arXiv:2603.21276*, 2026.

[55] J. Zhan, H. Shen, Z. Lin, and T. He, "Prism: Privacy-aware routing for adaptive cloud-edge llm inference via semantic sketch collaboration," *arXiv preprint arXiv:2511.22788*, 2025.

[56] Y. Lu, Z. Fang, P. Qiao, Z. Lin, J. Yang, Y. Zhang, P. L. Yee, Z. Chen, and J. Luo, "Conflict-aware federated fine-tuning of large language models with mixture-of-experts," *arXiv preprint arXiv:2606.15625*, 2026.

[57] Z. Lin, G. Qu, X. Chen, and K. Huang, "Split Learning in 6G Edge Networks," *IEEE Wirel. Commun.*, 2024.

[58] G. Qu, Q. Chen, W. Wei, Z. Lin, X. Chen, and K. Huang, "Mobile edge intelligence for large language models: A contemporary survey," *IEEE Communications Surveys & Tutorials*, 2025.

[59] M. Hu, J. Zhang, X. Wang, S. Liu, and Z. Lin, "Accelerating Federated Learning with Model Segmentation for Edge Networks," *IEEE Trans. Green Commun. Netw.*, 2024.

[60] S. Lyu, Z. Lin, G. Qu, X. Chen, X. Huang, and P. Li, "Optimal resource allocation for u-shaped parallel split learning," in *2023 IEEE Globecom Workshops (GC Wkshps)*, 2023, pp. 197–202.

[61] Y. Zhang, Z. Chen, X. Hu, J. Zhao, and Y. Gao, "S-leon: An efficient split learning framework over heterogeneous leo satellite networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 37, no. 1, pp. 106–121, 2025.

[62] W. Wei, Z. Lin, T. Li, X. Li, and X. Chen, "Pipelining split learning in multi-hop edge networks," *arXiv preprint arXiv:2505.04368*, 2025.

[63] Z. Lin, Y. Zhang, Z. Chen, Z. Fang, C. Wu, X. Chen, Y. Gao, and J. Luo, "LEO-Split: A Semi-Supervised Split Learning Framework over LEO Satellite Networks," *IEEE Trans. Mobile Comput.*, 2025.