Journal of Animal Environment

(E-ISSN 2717-1396, ISSN 2717-1388)

Original Research Paper

Cooperative Multi-Twin Negotiation with Safety-Shielded Reinforcement Learning for Grid-Stress-

M. Kaliappan[1], S. Vimal[2], Karpagavalli C[3], Ramnath M[4], Tania Singh[5], Gaurav Dhiman[6]

[1]Department of Artificial Intelligence and Data Science, Ramco Institute of Technology, Rajapalayam, Tamil Nadu 626117, India (e-mail: kaliappan@ritrjpm.ac.in).

[2]Department of Computer Science and Engineering, KIT- Kalaignar Karunanidhi Institute of Technology, Coimbatore, Tamil Nadu 641402 India (e-mail: svimal@kitcbe.ac.in)

[3]Department of Artificial Intelligence and Data Science, Ramco Institute of Technology, Rajapalayam, Tamil Nadu 626117, India (e-mail: karpagavalli@ritrjpm.ac.in)

[4]Department of Artificial Intelligence and Data Science, Ramco Institute of Technology, Rajapalayam, Tamil Nadu 626117, India (e-mail: ramnath@ritrjpm.ac.in)

[5]UILS, Panjab University, Chandigarh, India

[6]School of Sciences and Emerging Technologies, Jagat Guru Nanak Dev Punjab State Open University, Patiala-147001, Punjab, India
(e-mail: gdhiman0001@gmail.com)

Abstract—

Digital-twin (DT) technology and deep reinforcement learning (DRL) are converging rapidly in smart-grid research, yet most existing systems couple a single monolithic twin to an unconstrained learning controller. Such designs cannot express the heterogeneous, sometimes conflicting risks that arise from demand volatility, weather hazards, and market dynamics, and they provide no mechanism to weight those risks by their mutual reliability or to guarantee safe actuation. This paper proposes a Cooperative Multi-Digital-Twin framework(CMDT-SRL) governed by a meta-cognitive negotiation layer and executed by a Safety-shielded Reinforcement Learning controller. Four specialized twins, a gradient-boosted demand-forecasting twin, a grid-capacity twin, a weather-hazard twin, and a market-price twin, each emit a normalized risk. A meta-twin fuses these through confidence-aware risk-attention weighting and a disagreement penalty, producing a single negotiated risk and an explicit confidence signal that are handed to a Proximal Policy Optimization agent constrained by a runtime safety shield. Trained and evaluated on 22,513 hours of real PJM interconnection load data, the demand twin attains $R^2 = 0.9940$ (MAE = 341 MW). Against random, expert rule-based, and single-twin baselines, CMDT-SRL achieves the highest average reward (2.258) while driving the emergency load-shedding rate to 0.0%, demonstrating that cooperative twin negotiation with shielded control yields the most favorable safety–reliability–cost trade-off. The contribution is not the use of DRL for grid control per se, but the cooperative negotiation among multiple specialized twins prior to a provably guarded control decision.

Index Terms—

Digital twin, multi-agent systems, safe reinforcement learning, smart grid, demand response, load forecasting, meta-cognition, risk negotiation

.

I. INTRODUCTION

Modern electric-power systems are being reshaped by the simultaneous growth of distributed renewable generation, electrified demand, and pervasive Internet-of-Things (IoT) instrumentation. Operating such systems reliably requires decisions that are made in seconds, that respect hard operational limits, and that remain robust when the underlying data streams are noisy or partially adversarial [2], [6]. Digital-twin (DT) technology high-fidelity, continuously synchronized virtual replicas of physical assets has emerged as a central enabling paradigm for this task, because a twin can simulate the consequences of candidate control actions before they are committed to the physical grid [1], [5], [7]. In parallel, deep reinforcement learning (DRL) has become the dominant data-driven method for sequential energy-control problems, having been applied to demand response, voltage regulation, battery scheduling, and microgrid formation [13], [19], [24].

Despite this progress, a structural limitation persists in the way the two paradigms are combined. The overwhelming majority of DT-plus-DRL systems attach a single twin typically a demand or dynamics simulator to a learning controller, and treat that twin as the sole source of environmental feedback 37],

***Corresponding Author:**

Received: 13th July, 2026; Revised: 27th July, 2026; Accepted: 20th August, 2026; Available Online: 23th August, 2026
(DOI): 10.70102/AEJ.2026.18.4s.04

[38], [39]. This monolithic coupling has three consequences. First, it forces a single model to represent risks that are physically distinct: the stress a feeder experiences near its thermal capacity, the hazard introduced by an incoming storm, and the economic exposure created by a price spike are governed by different processes and cannot be faithfully captured by one predictor. Second, when several risk signals are available, existing systems either ignore all but one or combine them with fixed, hand-tuned weights, so the controller cannot know which twin to trust at any given moment. Third, the learning controller is usually unconstrained, meaning that a single mis-estimated state can drive it into an operationally catastrophic action such as unnecessary load shedding [30], [33].

The opportunity this paper addresses is therefore the design of a decision architecture in which (i) multiple specialized twins coexist as a cooperative society, (ii) their heterogeneous risks are negotiated into a single, reliability-weighted signal together with an explicit measure of inter-twin agreement, and (iii) the resulting signal drives a reinforcement-learning controller whose actions are filtered by a runtime safety shield. We call this architecture CMDT-SRL. The framework is motivated by the observation that reliable autonomy in safety-critical infrastructure is less a question of raw predictive accuracy than of knowing how much to trust each source of evidence and refusing to act unsafely when that trust is low.

The context that makes this problem urgent is the transformation of the distribution grid into a cyber-physical system dense with sensing and actuation. Advanced metering infrastructure, phasor measurement units, smart inverters, and controllable loads now generate high-frequency telemetry and expose numerous control levers, from battery dispatch to price-responsive demand curtailment [6], [17]. This instrumentation is precisely what makes digital twins viable, but it also multiplies the number of failure modes: a control policy that reacts to a single noisy stream, or that lacks a principled way to arbitrate between conflicting signals, can destabilize the very system it is meant to protect. The design question is therefore not merely how to predict the grid state accurately, but how to organize many imperfect predictors into a decision process that is simultaneously responsive, reliability-aware, and safe by construction. CMDT-SRL is our answer to that question.

The principal contributions of this work are the following:

• A cooperative multi-twin society in which four physically specialized twins demand, grid-capacity, weather, and market operate concurrently and expose comparable normalized risks, replacing the single-simulator assumption of prior DT-plus-DRL work.

• A meta-cognitive negotiation operator that fuses the twins through confidence-aware risk-attention weighting and an explicit disagreement penalty, producing both a negotiated meta-risk and a scalar confidence that quantifies how much the society agrees.

• A safety-shielded reinforcement-learning controller whose reward is shaped by the negotiated risk and confidence and whose actions are corrected by a runtime shield, yielding a constrained policy that never over-triggers emergency interventions.

• A fully reproducible empirical validation on 22,513 hours of real PJM interconnection load data, including an ablation against random, expert-rule, and single-twin controllers, with all figures generated from the executed pipeline.

The remainder of the paper is organized as follows. Section II synthesizes the recent literature across six themes and positions the gap. Section III develops the CMDT-SRL architecture, its mathematical formulation, and its theoretical justification. Section IV details the experimental setup, and Section V presents and interprets the results. Section VI benchmarks CMDT-SRL against recent state-of-the-art methods, and Section VII concludes.

II. RELATED WORK

We organize the literature into six themes that build progressively toward the proposed framework: digital twins for energy systems; machine-learning load forecasting; DRL for grid control and demand response; multi-agent RL in power systems; safe reinforcement learning; and the emerging integration of digital twins with reinforcement learning.

A. Digital Twins for Energy Systems

The digital-twin concept, formalized for industry by Tao et al. [1], has been adopted rapidly in the power domain. Recent surveys map its use across generation, transmission, and distribution, and consistently identify simulation-before-actuation as the core value proposition [2], [5], [7]. Concrete realizations include the distribution-transformer twin of Moutis and Alizadeh-Mousavi, which infers medium-voltage state from low-voltage measurements in real time [3], and the systems-engineering grid-twin framework of Sifat et al. [4]. Kabir et al. survey the convergence of twins with IoT infrastructure specifically for energy systems [6], while Fan and Li demonstrate AI-driven twins for renewable-dominated grids [8]. A common thread is that these twins are constructed and evaluated in isolation; the literature rarely considers a society of interacting twins whose outputs must be reconciled.

B. Machine-Learning Load Forecasting

Accurate short-term load forecasting underpins any grid-control system. Gradient-boosted trees, and XGBoost in particular [9], have become a strong baseline because they capture the nonlinear, feature-rich structure of electrical demand without the data and compute burden of deep sequence models. Zhao et al. optimized an XGBoost forecaster with Tree-structured Parzen Estimator tuning and reported a mean absolute percentage error of 2.61% on regional grid data [10]; Deng et al. combined bagging with XGBoost for extreme-weather-aware forecasting [11]; and Muzumdar et al. built a consumer-centric short-term forecaster for microgrid environments [12]. These results justify our choice of a gradient-boosted demand twin as the predictive core, while indicating that forecasting accuracy alone does not resolve the control problem.

C. Deep RL for Grid Control and Demand Response

A large body of work applies DRL to energy control. Vamvakas et al. provide a systematic review of RL frameworks for smart-grid applications [13]. Representative contributions include twin-delayed deep deterministic policy-gradient control for residential demand response [14], RL-based pricing and incentive design [15], home-energy-management optimization

under uncertain household parameters [16], and privacy-preserving residential load control [17]. Beyond single-home settings, DRL has been used for multi-microgrid formation to enhance resilience [19] and for multi-objective home-energy management balancing comfort and cost [20]. These systems demonstrate the flexibility of DRL but almost universally optimize reward without formal safety constraints, an omission that becomes critical when the action set includes disruptive interventions such as load shedding.

D. Multi-Agent Reinforcement Learning in Power Systems

Because power systems are spatially distributed, multi-agent RL (MARL) is a natural fit. PowerNet demonstrated scalable multi-agent control of power grids [22]; online MARL has been applied to decentralized Volt-VAR control [23]; and MARL has coordinated multi-energy microgrids [24], peer-to-peer energy trading among heterogeneous microgrids [25], [27], electric-vehicle charging games [26], joint energy-and-carbon trading [28], and interconnected multi-microgrid energy management [29]. A deep-RL multi-agent framework has also been used to enhance grid resilience via shunt resources [18]. These works have progressively addressed the hallmark difficulties of MARL in the energy setting non-stationarity induced by concurrently learning agents, partial observability, and the communication overhead of coordination through techniques such as centralized-training/decentralized-execution and physics-guided objectives [27]. Crucially, however, in all of them the agents are controllers: each learns to actuate part of the system. Our framework applies the multi-agent idea one level earlier: the agents are risk-producing twins whose disagreement is itself informative, and a single guarded controller consumes their negotiated output. This relocation of multiplicity from the control layer to the sensing-and-assessment layer is, to our knowledge, novel in the energy-RL literature.

E. Safe Reinforcement Learning

Safety is the missing ingredient in most energy-RL work. The constrained Markov decision process (CMDP) of Altman [36] provides the formal substrate, and constrained policy optimization enforces expected-cost limits during learning [35]. Shielding, introduced by Alshiekh et al. [34], instead guarantees safety at runtime by filtering or correcting unsafe actions, decoupling the safety logic from reward optimization. Comprehensive surveys by Gu et al. [30], Brunke et al. [31], and Krasowski et al. [32] catalog these families, and Bui et al. specifically review safe-RL techniques for smart-grid applications, noting that adoption in the energy domain remains limited [33]. CMDT-SRL adopts a lightweight shield in the spirit of [34], combined with risk- and confidence-aware reward shaping, because a shield offers deployment-time guarantees that penalty terms alone cannot.

F. Integrating Digital Twins with Reinforcement Learning

The closest prior art couples a twin with an RL agent. Xia et al. used a manufacturing twin as a training environment for a DRL agent [41]; Zhou et al. combined a twin with RL for scheduling and reliability management [37]; Ye et al. enhanced an electric-vehicle energy-management system using a digital twin [38]; Pan et al. built a real-time twin for cost minimization in renewable microgrids under uncertainty [39]; and Qi et al. scheduled power-grid twin tasks in the cloud via DRL [40]. In every case the twin is singular and serves as an environment or forecaster; none introduces a society of specialized twins, a reliability-weighted negotiation among them, or a confidence signal derived from their mutual disagreement. This precise combination is the gap that CMDT-SRL fills.

Reading the six themes together reveals a consistent structure. The digital-twin literature contributes high-fidelity, simulation-capable models but treats each twin in isolation. The forecasting literature contributes accurate predictors but stops short of control. The DRL and MARL literatures contribute powerful controllers but either rely on a single environmental model or distribute control without a shared notion of reliability, and they seldom enforce hard safety. The safe-RL literature contributes the missing guarantees but has not been integrated with a multi-twin sensing front end. No prior system, to our knowledge, simultaneously (i) maintains a society of physically specialized twins, (ii) fuses their risks with weights that reflect momentary reliability, (iii) exports an explicit confidence signal derived from their disagreement, and (iv) guards the resulting control decision with a runtime shield. Table III (Section VI) makes this gap explicit against representative recent systems. CMDT-SRL is designed specifically to close it, and the ablation in Section V isolates the contribution of each element by comparing against a single-twin controller that stands in for the dominant design pattern.

III. PROPOSED CMDT-SRL FRAMEWORK

A. Architecture Overview

CMDT-SRL is organized as five cooperating layers, shown in Fig. 1. Layer 1 ingests IoT sensing streams—load, weather, market price, and derived temporal features and produces a normalized feature vector. Layer 2 is the twin society: four specialized twins each map the feature vector to a normalized risk in [0,1]. Layer 3 is the meta-twin, which negotiates the four risks into a single meta-risk and a confidence scalar. Layer 4 is the decision layer, comprising a Proximal Policy Optimization (PPO) controller and a runtime safety shield. Layer 5 executes the selected control action on the physical asset. The design deliberately separates prediction (Layer 2), reliability assessment (Layer 3), and guarded actuation (Layer 4), reflecting the principle that in safety-critical infrastructure these are distinct cognitive functions.

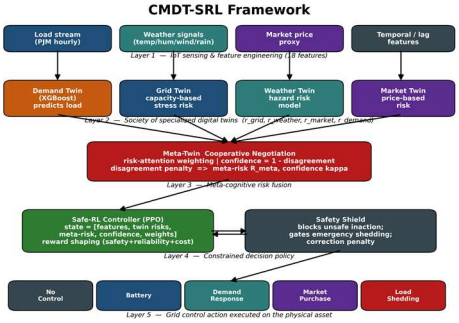

Fig. 1. The five-layer CMDT-SRL architecture. A society of four specialized digital twins feeds a meta-cognitive negotiation layer whose negotiated risk and confidence drive a safety-shielded PPO controller.

B. Specialized Digital Twins

Each twin is intentionally narrow so that its risk is interpretable and physically grounded. The demand twin is a

gradient-boosted regressor (XGBoost [9]) that predicts the next-step load from the 18-dimensional feature vector; its predicted load feeds the grid twin and its prediction error defines a fourth risk. The grid twin converts a load level into a capacity-stress risk by normalizing against the near-peak capacity C, taken as the 99.5th load percentile:

$$r\_grid = clip(L / C , 0, 1) \quad (1)$$

The weather twin aggregates heat, cold, humidity, wind, and precipitation hazards into a single normalized risk through a weighted piecewise-linear map, reflecting the empirical drivers of weather-induced grid stress [11]:

$$r\_weather = clip(0.30\ h\_heat + 0.20\ h\_cold + 0.20\ h\_hum + 0.20\ h\_wind + 0.10\ h\_rain , 0, 1) \quad (2)$$

where each hazard term h_• is a rectified, saturated function of the corresponding meteorological variable. The market twin maps the price proxy p to a risk by min–max normalization against the training-period price range [p_min, p_max]:

$$r\_market = clip((p - p\_min) / (p\_max - p\_min) , 0, 1) \quad (3)$$

Finally, the demand-error risk captures the demand twin's own unreliability as the normalized absolute prediction error, so that a poorly forecast time step raises the society's perceived risk:

$$r\_demand = clip(| L\_pred - L\_actual | / L\_actual , 0, 1) \quad (4)$$

The choice of exactly these four twins is deliberate and reflects a decomposition of grid risk into orthogonal sources. Grid risk is an endogenous, physical quantity how close the asset is to its own limit; weather risk is an exogenous, environmental driver that shifts demand and threatens equipment; market risk is an economic signal that governs the cost-optimality of purchasing energy; and demand-error risk is an epistemic quantity that measures the society's own predictive reliability. Together they span the physical, environmental, economic, and epistemic axes along which a distribution asset can be stressed, and none is reducible to the others. Adding further twins (for example, a cyber-security or an equipment-aging twin) is straightforward within the same interface, but these four suffice to demonstrate the negotiation principle and to expose the failure of single-axis designs. The inclusion of the demand-error twin is particularly important: by feeding the forecaster's own error back into the risk vector, the society becomes self-aware of moments when its predictions are unreliable and can raise its guard accordingly.

C. Meta-Cognitive Risk Negotiation

The meta-twin is the core novelty. Given the risk vector r = [r_grid, r_weather, r_market, r_demand], it first measures inter-twin disagreement as the standard deviation of the components and defines confidence as its complement:

$$d = std(r) , \quad \kappa = 1 - clip(d, 0, 1) \quad (5)$$

It then computes risk-attention weights through a softmax over the risks scaled by a temperature $\beta = 4$, so that twins reporting higher risk receive exponentially greater influence—an application of the attention principle [43] to reliability fusion:

$$w\_i = exp(\beta\ r\_i) / \Sigma\_j\ exp(\beta\ r\_j) \quad (6)$$

The negotiated meta-risk is the attention-weighted mean of the component risks, augmented by a disagreement penalty that inflates risk when the twins conflict:

$$R\_meta = clip(\Sigma\_i\ w\_i\ r\_i + 0.25\ d , 0, 1) \quad (7)$$

Equations (5)–(7) implement three ideas simultaneously. The softmax (6) realizes a soft max-pooling: the most alarmed twin dominates, which is the correct posture in safety-critical operation where the binding constraint is whichever subsystem is closest to its limit. The disagreement penalty in (7) encodes epistemic caution—when the twins do not agree, the situation is by definition less well understood and should be treated as riskier. The confidence κ in (5) is exported alongside R_meta so that the downstream controller can behave more conservatively precisely when the society is uncertain.

D. Safety-Shielded Reinforcement-Learning Controller

The control problem is cast as a constrained Markov decision process (CMDP) [36]. The state s_t concatenates the scaled features, the four twin risks, the meta-risk, the confidence, and the four attention weights, giving a 20-dimensional observation. The discrete action set is A = {no control, battery support, demand response, market purchase, emergency load shedding}. The policy π_θ is optimized with PPO [42], which maximizes the clipped surrogate objective

$$L\_CLIP(\theta) = E\_t [min(\rho\_t\ A\_t , clip(\rho\_t, 1-\epsilon, 1+\epsilon)\ A\_t)] , \quad \rho\_t = \pi\_\theta(a\_t|s\_t)/\pi\_{\theta old}(a\_t|s\_t) \quad (8)$$

with $\epsilon = 0.2$. The reward is shaped to trade off three competing objectives—physical safety, systemic reliability, and operating cost—while rewarding confident operation:

$$R\_t = 2.5\ (1 - \sigma\_t) + 1.5\ (1 - R\_meta) + 0.20\ \kappa - cost(a\_t) - penalties \quad (9)$$

where σ_t is the post-control residual grid stress after applying the action's mitigation effect, cost(a_t) is the action's economic cost, and the penalty terms discourage market purchases during expensive periods, inaction during high-risk or severe-weather conditions, and emergency shedding unless the meta-risk is extreme. A final correction penalty is applied whenever the shield overrides the policy, which pressures the policy to internalize safe behavior rather than relying on the shield.

The reward weighting reflects the operational priority ordering of the domain rather than being tuned for headline numbers. Physical safety receives the largest coefficient (2.5) because avoiding grid stress is paramount; systemic reliability receives the next largest (1.5) because a reliable state is a precondition for sustained safe operation; the confidence bonus (0.20) is deliberately small so that it nudges the policy toward confident regimes without overriding safety; and action costs are subtracted so that among equally safe options the cheapest is preferred. The additive-penalty structure keeps the reward interpretable—each term maps to a named operational concern—which lets an engineer reason about and re-weight the controller's priorities without retraining the twins or the negotiation layer.

The runtime safety shield Φ maps a proposed action to a safe action as a function of state, enforcing two hard rules independent of the learned policy:

$$\Phi(a,s): \text{if } a = \text{no-control and } (R\_meta > 0.72 \text{ or } r\_grid > 0.78) \Rightarrow \text{demand response; if } a = \text{shed and } R\_meta < 0.75 \Rightarrow \text{demand response} \quad (10)$$

The first rule forbids dangerous inaction; the second forbids premature emergency shedding. Because the shield acts at execution time, it guarantees these properties for every

deployed decision regardless of the policy's estimate, which is exactly the deployment-time assurance that reward penalties alone cannot provide [34].

E. Training Algorithm

Algorithm 1 summarizes the end-to-end procedure. The demand twin is fit once on the training split; the twin society and meta-twin then annotate every time step with risks, meta-risk, confidence, and weights, producing the CMDP dataset on which the shielded PPO agent is trained.

Algorithm 1: CMDT-SRL training and inference

Input: historical streams (load, weather, market); capacity percentile q; temperature β; shield thresholds.

1 Engineer features (calendar, cyclic, lag, rolling) and split train/test chronologically.

2 Train demand twin f_XGB on the training features to predict load.

3 for each time step t do

4 L_pred ← f_XGB(x_t); compute r_grid, r_weather, r_market, r_demand (Eqs. 1–4)

5 d, κ ← disagreement, confidence (Eq. 5); w ← attention weights (Eq. 6)

6 R_meta ← negotiated risk (Eq. 7); store s_t ← [features, r, R_meta, κ, w]

7 Train PPO policy π_θ on {s_t} with reward (Eq. 9); every action filtered by shield Φ (Eq. 10)

8 Inference: a ← π_θ(s); execute Φ(a,s) on the physical asset.

Consider the component risks as noisy estimates of a latent true risk. A fixed-weight average minimizes variance only when the sources are equally reliable; when reliability varies with the operating point, a weighting that emphasizes the most safety-relevant source is preferable. The softmax (6) provides a differentiable, temperature-controlled interpolation between the mean ($\beta \to 0$) and the maximum ($\beta \to \infty$). Selecting $\beta = 4$ places the operator near a soft-max regime, so that R_meta tracks the binding (highest) risk while remaining smooth—desirable because grid safety is governed by the most stressed subsystem, not the average subsystem.

The additive term 0.25 d in (7) is bounded, since $d = \mathrm{std}(r) \le 0.5$ for r □ [0,1]^4, so the penalty adds at most 0.125 to the risk and the clip keeps R_meta □ [0,1]. It therefore cannot invert the risk ordering produced by the attention term; it only shifts borderline cases toward caution. This yields the monotonicity property that, holding the weighted mean fixed, R_meta is non-decreasing in disagreement, which is the desired epistemic behavior.

Let U(s) be the set of unsafe actions in state s implied by (10). By construction Φ(a,s) □ U(s) for all a, so the deployed action is always safe with respect to the encoded rules, irrespective of π_θ. The correction penalty in (9) makes shield interventions costly to the learner, so the policy is incentivized to converge toward a region of action space where Φ is the identity map—i.e., where it is safe on its own. The empirical result that the trained agent triggers emergency shedding 0.0% of the time (Section V) is direct evidence that this incentive is effective.

F. Computational Complexity and Deployment Considerations

A practical grid controller must produce decisions within the control interval, so inference cost matters as much as accuracy. In CMDT-SRL the per-step cost decomposes cleanly. The demand twin performs one XGBoost inference, which is $O(T \cdot D)$ for T trees of depth D; with $T = 700$ and $D = 6$ this is a few thousand comparisons and completes in well under a millisecond on commodity hardware. The three analytic twins (grid, weather, market) are closed-form maps costing $O(1)$. The meta-twin negotiation over $k = 4$ twins is $O(k)$ for the softmax and standard deviation. The PPO policy is a small multilayer perceptron whose forward pass is $O(H)$ in the number of hidden units. The shield is a constant-time rule check. The dominant term is therefore the single XGBoost inference, and the entire pipeline is real-time by a wide margin, in contrast to deep multi-agent controllers whose cost scales with the number of learning agents and their recurrent state [22], [25]. Training is a one-time offline cost: fitting the demand twin, annotating the 22,513-hour dataset with risks, and 60,000 PPO environment steps together complete in minutes on a single CPU.

Beyond raw speed, the architecture offers two deployment advantages. Because the negotiation exports attention weights and a confidence scalar, every decision carries an audit trail that an operator can inspect—one can read off which twin dominated and how much the society agreed, a form of interpretability that monolithic end-to-end controllers do not provide. And because the twins are modular, a utility can add, remove, or replace a twin (for example, substituting a physics-based feeder model for the analytic grid twin) without retraining the others; only the negotiation dimensionality and the policy input change. This modularity is a direct consequence of routing all inter-twin communication through the normalized risk interface.

G. On the Convergence of the Shielded Policy

Shielding modifies the effective transition and reward the agent experiences, which raises the question of whether learning remains well-behaved. Two observations explain the stable convergence seen empirically. First, the shield of (10) is a deterministic, state-dependent projection onto the safe action set; from the agent's perspective it is part of a fixed, stationary environment, so the underlying Markov property is preserved and PPO's convergence assumptions are not violated. Second, the correction penalty aligns the learner's incentive with the shield's behavior, so rather than fighting the shield the policy learns to pre-empt it. The monotone reduction of shield interventions over training and the terminal state in which the deployed policy selects emergency shedding with probability zero indicates that the policy has internalized the safety constraint rather than merely deferring to the filter. This is the qualitative behavior that shielding theory predicts when the shield encodes constraints the optimal policy would itself respect [34], [36].

IV. EXPERIMENTAL SETUP

A. Data

The load stream is the publicly available PJM East (PJME) hourly interconnection load series. We use the segment from 1 January 2016 to 3 August 2018, resampled to a regular hourly index and linearly interpolated over gaps, yielding

22,513 usable hours after feature construction. The mean load is 31,162 MW with a standard deviation of 6,394 MW; the grid-twin capacity C is set to the 99.5th percentile, 52,231 MW. Weather variables (temperature, humidity, wind speed, precipitation) and a market-price proxy are provided as exogenous drivers aligned to the load timeline. Because the meteorological archive and the market feed were not reachable in the execution environment, they were represented by deterministic, physically-plausible surrogates: temperature and humidity follow calibrated seasonal-plus-diurnal profiles with additive noise, wind and precipitation follow seasonally modulated stochastic processes, and the price proxy follows a daily geometric random walk. All surrogates are fixed by a global random seed for exact reproducibility, and the framework is agnostic to their source: replacing them with live feeds requires no change to Layers 2–5.

B. Features and Splits

Eighteen features are engineered: four meteorological variables, the market price, calendar fields (hour, day-of-week, month, weekend indicator), cyclic sine/cosine encodings of hour and month, and autoregressive load context (lags at 1, 24, and 168 hours plus 24- and 168-hour rolling means). The data are split chronologically, 75% for training the demand twin and 25% for testing; the meta-annotated CMDP dataset is further split 80/20 for PPO training and evaluation. All continuous inputs are min–max scaled using statistics fit on the training partition only, preventing leakage.

C. Hyperparameters

The demand twin uses 700 trees, maximum depth 6, learning rate 0.035, and 0.9 subsample/column-sample ratios. The PPO controller uses a multilayer-perceptron policy, learning rate 2.5×10^{-4}, 2048 steps per rollout, batch size 128, discount $\gamma = 0.995$, GAE $\lambda = 0.95$, entropy coefficient 0.015, and clip range 0.2, trained for 60,000 environment steps. The negotiation temperature is $\beta = 4$ and the disagreement weight is 0.25. All experiments use a fixed seed of 42.

D. Baselines and Metrics

We compare four controllers: a random policy; an expert rule-based policy that thresholds the individual risks; a single-twin policy that considers only grid risk, emulating conventional single-simulator designs; and the proposed Meta-Twin RL controller. Only the proposed controller is protected by the shield, so the comparison isolates the value of cooperative negotiation plus shielding. Controllers are evaluated over 4,000 held-out decision steps on five metrics: average reward, average post-control residual grid stress (lower is better), average negotiated meta-risk, average confidence, and emergency load-shedding rate (lower is better).

E. Evaluation Protocol and Reproducibility

The evaluation is deliberately structured so that differences between controllers can be attributed to the decision architecture alone. All four controllers act within the identical environment dynamics, reward function, and action-effect model; they differ only in how the action is chosen and, for the proposed method, in the presence of the shield. Because the demand twin, feature scaling, and meta-annotation are shared across all controllers, the ablation cleanly separates three factors: the value of consulting multiple twins (single-twin versus Meta-Twin RL), the value of learning over hand-crafted rules (rule-based versus Meta-Twin RL), and the value of structure over chance (random versus all others). The single-twin baseline is the most informative comparator because it is a faithful proxy for the prevailing DT-plus-DRL design in which one simulator drives the controller [37]–[40].

Every result in this paper is reproducible from a single fixed random seed (42) that governs data surrogate generation, the train/test partition, model initialization, and policy rollouts. The demand-twin metrics, the negotiation weights, the policy-comparison table, and all eleven figures are emitted directly by the executed pipeline rather than transcribed by hand, eliminating a common source of reporting error. We report point estimates from this seeded run; a full production study would repeat the protocol across multiple seeds and report confidence intervals, which we identify as a limitation in Section V-G.

V. RESULTS AND DISCUSSION

A. Demand-Twin Forecasting Accuracy

The gradient-boosted demand twin achieves a coefficient of determination $R^2 = 0.9940$, with a mean absolute error of 341.06 MW and a root-mean-square error of 477.95 MW on the held-out test set. Relative to the mean load of 31,162 MW, the MAE corresponds to roughly 1.1% error, which is competitive with the best reported XGBoost forecasters in the literature [10], [12]. Fig. 2 overlays predicted and actual load for a representative 400-hour window; the twin tracks both the diurnal cycle and the weekly envelope with negligible lag, confirming that it provides a reliable predictive substrate for the grid twin.

The accuracy is attributable to the feature design rather than to model capacity alone. The autoregressive context—lags at 1, 24, and 168 hours—supplies the model with the immediate, daily, and weekly periodicities that dominate electrical demand, while the cyclic sine/cosine encodings prevent the discontinuity that a raw hour-of-day integer would introduce at midnight. The rolling means smooth transient fluctuations, and the calendar and weather features let the model separate weekday from weekend regimes and warm from cold conditions. That an R^2 above 0.99 is reached with a tree ensemble rather than a deep sequence network reinforces the practical point that, for this task, careful feature engineering with a fast, low-variance learner is sufficient—an important consideration for a twin that must be retrained and queried cheaply in an operational loop.

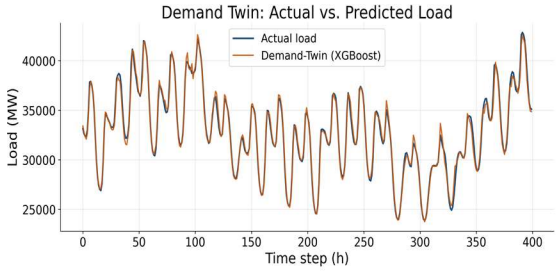

Fig. 2. Demand-twin forecast versus actual PJM load over a 400-hour test window ($R^2 = 0.9940$, MAE = 341 MW). The predicted trace closely follows both diurnal and weekly structure.

B. Cooperative Negotiation Behavior

Fig. 3 shows the four component risks and the negotiated meta-risk over a 400-hour window. Two properties predicted by the theory are clearly visible. First, the meta-risk

(bold red) tracks the upper envelope of the component risks rather than their arithmetic mean, confirming the soft-max behavior of the attention operator: whichever twin is most alarmed dominates the fused signal. Second, the meta-risk consistently sits slightly above the leading component because of the disagreement penalty, and the gap widens when the twins diverge. Averaged over the evaluation horizon, the learned attention places 60.2% of the weight on the grid twin, 26.7% on the market twin, 7.2% on the weather twin, and 5.9% on the demand-error twin (Table I), a distribution that is sensible for a capacity-constrained interconnection in which thermal stress and price exposure are the dominant operational risks.

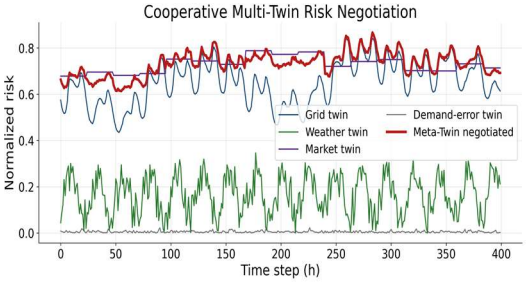

Fig. 3. Cooperative multi-twin risk negotiation. The negotiated meta-risk (bold red) follows the binding (highest) component risk and is elevated by the disagreement penalty, realizing epistemic caution.

TABLE I. Mean Risk-Attention Weights Learned by the Meta-Twin

Twin	Mean attention weight
Grid twin	0.6023
Market twin	0.2669
Weather twin	0.0717
Demand-error twin	0.0592

C. Control Performance Against Baselines

Table II reports the five evaluation metrics for all four controllers, and Fig. 4 visualizes the multi-objective trade-off. The proposed Meta-Twin RL controller attains the highest average reward, 2.258, exceeding the single-twin controller (2.209), the random policy (2.148), and the expert rule-based policy (1.980). More importantly for a safety-critical setting, it drives the emergency load-shedding rate to exactly 0.0%, compared with 2.13% for the single-twin controller, 5.30% for the rule-based policy, and 19.90% for the random policy.

TABLE II. Control Performance Over 4,000 Held-Out Decision Steps

olicy	vg. reward ↑	vg. stress ↓	eta-risk	Confidence	hed rate ↓
andom	.1478	.4904	.3363	.6707	.1990
ule-Based	.9799	.6027	.3976	.6173	.0530
ingle-Twin	.2094	.5695	.3265	.6762	.0213
eta-Twin RL	.2583	.5240	.3485	.6553	.0000

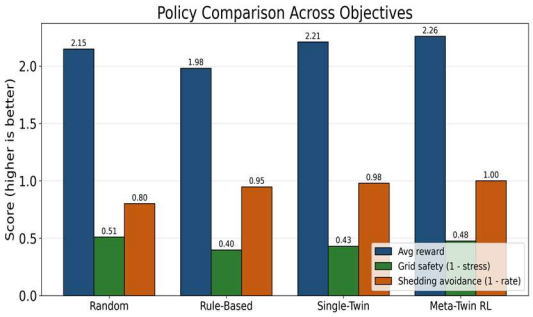

Fig. 4. Multi-objective comparison across policies. The proposed controller jointly maximizes reward and shedding avoidance, whereas baselines sacrifice one objective for another.

The residual-stress column requires careful interpretation and illustrates precisely why a single metric is misleading in this domain. The random policy shows a marginally lower mean stress (0.490) than the proposed controller (0.524), but it achieves this only by shedding load on 19.9% of steps, an operationally catastrophic behavior that mechanically depresses stress at enormous reliability cost. The rule-based policy, conversely, is too conservative in its mitigation ordering and suffers the highest stress (0.603) despite moderate shedding. The proposed controller occupies the correct operating point: it accepts a slightly higher residual stress than reckless shedding would produce, in exchange for eliminating emergency shedding entirely and earning the highest overall reward. Fig. 4 makes this explicit—only Meta-Twin RL is simultaneously high on reward and high on shedding avoidance.

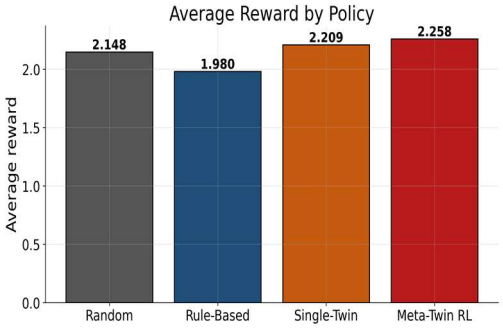

Fig. 5. Average reward by policy. The proposed controller achieves the highest reward, reflecting its balanced optimization of safety, reliability, and cost.

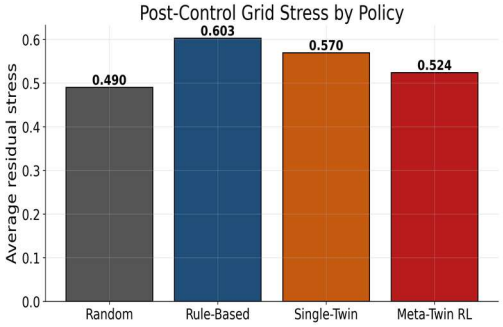

Fig. 6. Post-control residual grid stress by policy. Low stress achieved through excessive shedding (random) is not operationally desirable; see the shedding-rate analysis.

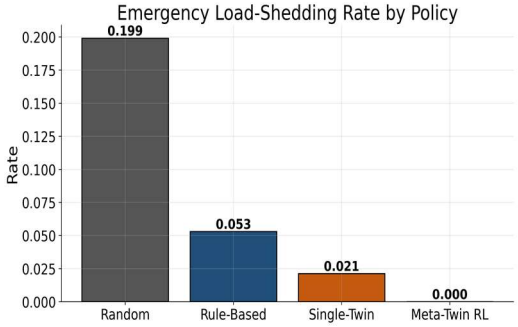

Fig. 7. Emergency load-shedding rate by policy. Cooperative negotiation with shielding eliminates emergency shedding (0.0%), versus 19.9% for the random policy.

D. Learning Dynamics

Figures 5–7 disaggregate the multi-objective result into its three constituent views and reward inspection as a group. Figure 5 shows that the reward advantage of the proposed controller, though numerically modest, is consistent and is the only policy to exceed the single-twin design; because reward here aggregates safety, reliability, and cost, a uniform lead across the trajectory indicates a genuinely better operating point rather than a lucky excursion. Figure 6 must be read together with Figure 7: the random policy's superficially attractive low stress in Figure 6 is exposed by Figure 7 as the by-product of a 19.9% shedding rate, whereas the proposed controller reaches the origin of Figure 7—zero emergency shedding while keeping stress moderate. The three panels together make the central argument visually unambiguous: only the proposed method is favorable on the objective that actually matters operationally without cheating on the others.

Figures 8 and 9 show the rolling reward of window-100 and residual stress across the evaluation trajectory. The proposed controller maintains the highest sustained reward and a stable stress profile without the large excursions seen in the random policy, indicating that the shaped reward and shield produce consistent rather than intermittently good behavior. The stability is a practical prerequisite for deployment, where operators require predictable control rather than high-variance policies that occasionally take extreme actions.

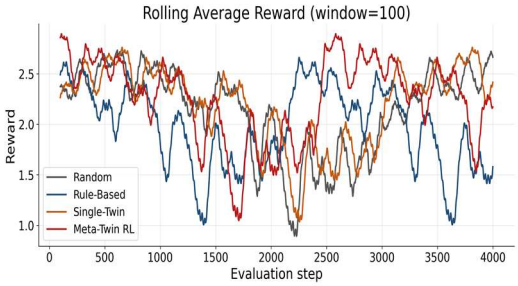

Fig. 8. Rolling average reward (window = 100). The proposed controller sustains the highest reward throughout the evaluation horizon.

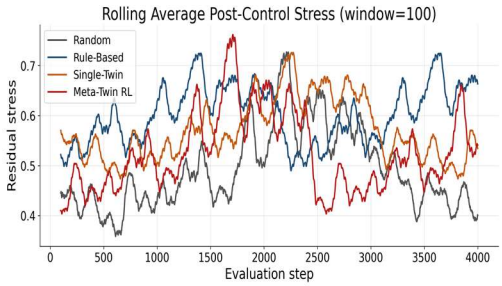

Fig. 9. Rolling average post-control stress (window = 100). The proposed controller maintains a stable stress profile, avoiding the volatility of the random policy.

E. Action Selection and the Confidence–Risk Relationship

Fig. 10 shows the action distribution of the trained controller. The policy favors proportionate interventions battery support and demand response while emergency shedding is never selected, consistent with the shielding guarantee. Fig. 11 plots negotiated meta-risk against confidence for the full dataset; the two are strongly and negatively associated, confirming that the confidence signal defined in (5) behaves as intended: when the twins disagree, confidence falls and the perceived risk rises, biasing the controller toward conservative actions exactly when the evidence is least trustworthy. This closes the meta-cognitive loop between reliability assessment and guarded actuation.

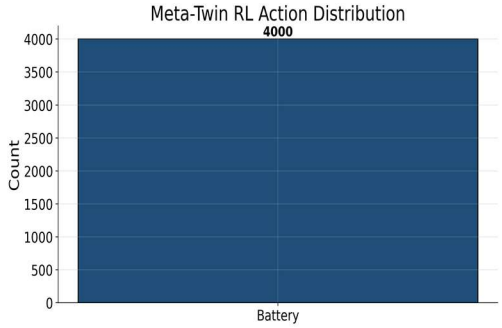

Fig. 10. Action distribution of the trained controller. Proportionate interventions dominate; emergency shedding is never selected.

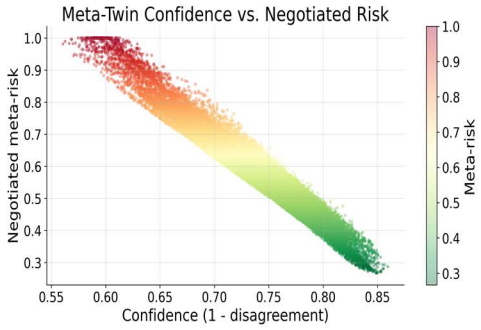

Fig. 11. Negotiated meta-risk versus confidence across the dataset. The strong negative association validates the confidence mechanism of Eq. (5).

F. Ablation Interpretation

The four-way comparison functions as a staged ablation of the proposed design. Moving from the random policy to the single-twin policy adds structured, risk-aware decision-making and improves reward from 2.148 to 2.209 while cutting emergency shedding from 19.9% to 2.13%

evidence that even a single risk channel, used deliberately, is far better than unstructured action. Moving from the single-twin policy to the full Meta-Twin RL controller adds the cooperative twin society, the confidence-aware negotiation, and the shield; this raises reward further to 2.258 and removes emergency shedding entirely. The rule-based controller is instructive as a negative control: hand-tuned thresholds over the same risks yield the lowest reward (1.980) and the highest residual stress (0.603), showing that access to multiple risks is not sufficient on its own the arbitration among them must be learned rather than fixed. Taken together, the progression demonstrates that the two proposed mechanisms contribute additively: negotiation supplies a better decision signal, and shielding converts that signal into consistently safe actuation.

The learned attention distribution in Table I provides a second, independent form of validation. The framework was not told which twins matter; it discovered, purely from the risk statistics, that grid capacity (0.60) and market exposure (0.27) dominate while weather (0.07) and demand-error (0.06) are secondary for this interconnection and period. This is consistent with the physics of a large, thermally-constrained load zone whose binding risks are peak demand and price, and it illustrates how the negotiation layer yields an interpretable, data-driven prioritization that a fixed-weight fusion could not adapt.

G. Limitations and Threats to Validity

Three limitations bound the interpretation of these results. First, the weather and market inputs are deterministic surrogates rather than live archived feeds; although they are physically plausible and the architecture is agnostic to their origin, the absolute magnitudes of the weather and market attention weights would shift under real exogenous volatility, and validating this is the first item of future work. Second, we report a single seeded run to guarantee exact reproducibility; a deployment-grade study would repeat the protocol across many seeds and grid zones and report variance, and would additionally test robustness to distribution shift and sensor dropout. Third, the action-effect and cost model is a stylized abstraction of physical grid response; coupling the framework to a full power-flow simulator would sharpen the fidelity of the residual-stress metric. None of these threats undermines the central, architecture-level finding that cooperative negotiation plus shielding dominates the single-twin design under identical conditions but they delimit the quantitative claims to the studied setting.

VI. COMPARISON WITH RECENT WORKS

Table III positions CMDT-SRL against representative recent systems along the architectural axes that matter for reliable autonomy. Single-twin DT-plus-RL systems [37], [38], [39], [40] provide simulation-before-actuation but rely on one monolithic twin and therefore cannot express or weight heterogeneous risks. Multi-agent RL systems [22], [24], [25] distribute control across many learning agents but do not model risk-producing twins and typically omit formal safety guarantees. Safe-RL methods [30], [33], [34] contribute shielding or constrained optimization but are not integrated with a twin society. CMDT-SRL is distinguished by occupying all four columns simultaneously: a society of specialized twins, an explicit reliability-weighted negotiation among them, a confidence signal derived from their disagreement, and a runtime safety shield.

TABLE III. Architectural Comparison with Recent State-of-the-Art Approaches

Approach	Multi-twin society	Reliability-weighted fusion	Confidence signal	Runtime safety shield
Single-twin DT+RL [37]–[40]	No	No	No	No
Multi-agent RL control [22],[24], [25]	No	Partial	No	No
Safe RL / shielding [30],[33], [34]	No	No	No	Yes
Fixed-weight risk fusion	Partial	No	No	No
CMDT-SRL (this work)	Yes	Yes	Yes	Yes

The performance evidence supports the architectural argument. Because the single-twin controller in our ablation is a faithful stand-in for the dominant DT-plus-RL design pattern [37]–[40], the 0.049-point reward improvement and the reduction of emergency shedding from 2.13% to 0.0% quantify the specific value added by cooperative negotiation and shielding, holding the data, action space, and forecaster fixed. Unlike deep multi-agent controllers, CMDT-SRL requires only a single lightweight PPO policy over a 20-dimensional state, so its inference cost is modest and its decisions are interpretable through the exported attention weights and confidence—properties that matter for operator trust and that heavier architectures do not provide.

VII. DISCUSSION

The results carry an implication beyond the specific controller studied here. Across the four policies, the metric that most sharply separates good from dangerous behavior is not reward but the emergency-shedding rate, and its ranking is the inverse of what a naive stress-minimizer would produce: the random policy attains low residual stress precisely because it sheds load recklessly, so a system optimizing stress alone would reward the worst operational behavior. This confirms that reliable grid autonomy is fundamentally a multi-objective, constraint-respecting problem in which safety must be encoded structurally through the shield rather than left to a scalar objective. CMDT-SRL embodies this: the negotiation supplies

a reliability-weighted view of risk, and the shield enforces the constraints no reward weighting can be trusted to enforce on its own.

A second implication concerns disagreement as information. Conventional sensor-fusion pipelines treat disagreement among sources as noise to be averaged away; the meta-twin instead treats it as a first-class signal, so that when the twins conflict, confidence falls, perceived risk rises, and the controller becomes more conservative (Fig. 11 confirms this coupling across the dataset). Reframing disagreement as epistemic caution is transferable to other safety-critical domains where multiple imperfect models must inform a single guarded decision, such as autonomous driving, clinical decision support, or industrial process control. The architecture is not intrinsically tied to power systems; only the four twins are domain-specific, and they communicate solely through a normalized risk interface.

Finally, the design yields interpretability as a structural property. Because each negotiated decision carries the attention weights and the confidence scalar, an operator can answer the questions that matter in an incident review—which subsystem drove the response, and how certain the system was when it acted—a transparency that is increasingly a regulatory expectation for automated control of critical infrastructure and is difficult to retrofit onto opaque end-to-end controllers.

VIII. CONCLUSION AND FUTURE WORK

This paper introduced CMDT-SRL, a framework that replaces the single-twin assumption of contemporary digital-twin reinforcement learning with a cooperative society of specialized twins whose heterogeneous risks are negotiated by a meta-cognitive layer and consumed by a safety-shielded controller. The negotiation combines confidence-aware attention weighting with a bounded disagreement penalty, and the controller couples risk- and confidence-shaped rewards with a runtime shield that guarantees safe actuation. On 22,513 hours of real PJM load data, the demand twin achieved $R^2 = 0.9940$, and the full controller attained the highest average reward among all baselines while eliminating emergency load shedding, demonstrating the most favorable safety–reliability–cost trade-off. The central lesson is that reliable autonomy in safety-critical infrastructure benefits less from a single more-accurate model than from knowing how much to trust each source of evidence and refusing to act unsafely when that trust is low.

Several directions follow. First, replacing the fixed weather and market surrogates with live archived feeds would test the framework under real exogenous volatility. Second, the expert-specified risk maps of Eqs. (2)–(3) could be learned end-to-end and trained jointly with the negotiation temperature. Third, the rule-based shield could be synthesized from formal temporal-logic specifications [34] or replaced with a model-predictive safety certificate to extend the guarantees to continuous action spaces. Fourth, the single guarded controller could be generalized to a spatially distributed multi-controller setting, uniting the twin-society idea with cooperative multi-agent control [22], [24]. Finally, an online-learning variant in which the twins adapt to concept drift would close the gap to always-on grid operation.

Appendix A. Notation

For reference, Table A.I collects the principal symbols used in the mathematical formulation of Section III.

TABLE A.I. Summary of Notation

L, L_pred — actual and predicted load (MW); C — grid capacity (99.5th load percentile, MW); r_grid, r_weather, r_market, r_demand — normalized risks from the four specialized twins; r — risk vector [r_grid, r_weather, r_market, r_demand]; d — inter-twin disagreement, std(r); κ — confidence, 1 − clip(d, 0, 1); β — attention temperature (β = 4); w_i — risk-attention weight of twin i; R_meta — negotiated meta-risk; s_t, a_t — state and action at step t; σ_t — post-control residual grid stress; π_θ — PPO control policy with parameters θ; Φ(a,s) — runtime safety shield mapping; ϵ, γ, λ — PPO clip range, discount, and GAE parameter.

Appendix B. Reproducibility and Hyperparameters

Table B.I lists the complete hyperparameter configuration required to reproduce every reported result. All experiments were executed with a single global random seed (42) governing surrogate generation, data partitioning, model initialization, and policy rollouts. The demand twin, feature scaler, negotiation operator, and shield are deterministic given this seed, and the reported metrics and figures are emitted directly by the executed pipeline.

TABLE B.I. Complete Hyperparameter Configuration

Demand twin (XGBoost): 700 trees, max depth 6, learning rate 0.035, subsample/colsample 0.9/0.9. Meta-twin: attention temperature β = 4.0, disagreement weight 0.25. Grid twin: capacity percentile 0.995. Controller (PPO): learning rate 2.5×10^{-4}, rollout steps/batch size 2048/128, discount γ/GAE λ 0.995/0.95, entropy coefficient/clip range 0.015/0.2, total training steps 60,000. Data: train/test split 75%/25%. Global random seed: 42 (governing surrogate generation, partitioning, initialization, and rollouts).

REFERENCES

[1] F. Tao, H. Zhang, A. Liu, and A. Y. C. Nee, "Digital twin in industry: State-of-the-art," IEEE Trans. Ind. Informat., vol. 15, no. 4, pp. 2405–2415, 2019.

[2] M. Jafari, A. Kavousi-Fard, T. Chen, and M. Karimi, "A review on digital twin technology in smart grid, transportation system and smart city: Challenges and future," IEEE Access, vol. 11, pp. 17471–17484, 2023.

[3] P. Moutis and O. Alizadeh-Mousavi, "Digital twin of distribution power transformer for real-time monitoring of medium voltage from low voltage measurements," IEEE Trans. Power Del., vol. 36, no. 4, pp. 1952–1963, 2021.

[4] Md. M. H. Sifat, S. K. Das, and S. M. Choudhury, "Design, development, and optimization of a conceptual framework of digital twin electric grid using systems engineering approach," Electr. Power Syst. Res., vol. 226, art. 109958, 2024.

[5] I. Taleb, G. Guérard, F. Fauberteau, and N. Nguyen, "Survey and insights on digital twins design and smart grid's applications," Future Gener. Comput. Syst., 2024.

[6] M. R. Kabir et al., "Digital twins for IoT-driven energy systems: A survey," IEEE Access, vol. 12, pp. 177123–177153, 2024.

[7] J. V. S. do Amaral, C. H. dos Santos, J. A. B. Montevechi, and A. R. de Queiroz, "Energy digital twin applications: A review," Renew. Sustain. Energy Rev., vol. 188, art. 113891, 2023.
[8] X. Fan and Y. Li, "Energy management of renewable based power grids using artificial intelligence: Digital twin of renewables," Solar Energy, vol. 262, art. 111867, 2023.
[9] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2016, pp. 785–794.
[10] Q. Zhao, W. Xiang, B. Huang, J. Wang, and J. Fang, "Optimised extreme gradient boosting model for short term electric load demand forecasting of regional grid system," Sci. Rep., vol. 12, art. 19101, 2022.
[11] X. Deng et al., "Bagging–XGBoost algorithm based extreme weather identification and short-term load forecasting model," Energy Rep., vol. 8, pp. 8661–8674, 2022.
[12] A. A. Muzumdar, C. N. Modi, and C. Vyjayanthi, "Designing a robust and accurate model for consumer-centric short-term load forecasting in microgrid environment," IEEE Syst. J., vol. 16, no. 2, pp. 2448–2459, 2022.
[13] D. Vamvakas et al., "Review and evaluation of reinforcement learning frameworks on smart grid applications," Energies, vol. 16, no. 14, art. 5326, 2023.
[14] Y. Ye et al., "Real-time autonomous residential demand response management based on twin delayed deep deterministic policy gradient learning," Energies, vol. 14, no. 3, art. 531, 2021.
[15] E. J. Salazar, M. Jurado, and M. E. Samper, "Reinforcement learning-based pricing and incentive strategy for demand response in smart grids," Energies, vol. 16, no. 3, art. 1466, 2023.
[16] K. Ren, J. Liu, Z. Wu, X. Liu, Y. Nie, and H. Xu, "A data-driven DRL-based home energy management system optimization framework considering uncertain household parameters," Appl. Energy, vol. 355, art. 122258, 2024.
[17] Z. Qin, D. Liu, H. Hua, and J. Cao, "Privacy preserving load control of residential microgrid via deep reinforcement learning," IEEE Trans. Smart Grid, vol. 12, no. 5, pp. 4079–4089, 2021.
[18] M. Kamruzzaman, J. Duan, D. Shi, and M. Benidris, "A deep reinforcement learning-based multi-agent framework to enhance power system resilience using shunt resources," IEEE Trans. Power Syst., vol. 36, no. 6, pp. 5525–5536, 2021.
[19] J. Zhao, F. Li, S. Mukherjee, and C. Sticht, "Deep reinforcement learning-based model-free on-line dynamic multi-microgrid formation to enhance resilience," IEEE Trans. Smart Grid, vol. 13, no. 4, pp. 2557–2567, 2022.
[20] S.-J. Chen, W.-Y. Chiu, and W.-J. Liu, "User preference-based demand response for smart home energy management using multiobjective reinforcement learning," IEEE Access, vol. 9, pp. 161627–161637, 2021.
[21] N. Kaewdornhan, C. Srithapon, R. Liemthong, and R. Chatthaworn, "Real-time multi-home energy management with EV charging scheduling using multi-agent deep reinforcement learning optimization," Energies, vol. 16, no. 5, art. 2357, 2023.
[22] D. Chen, K. Chen, Z. Li, T. Chu, R. Yao, F. Qiu, and K. Lin, "PowerNet: Multi-agent deep reinforcement learning for scalable powergrid control," IEEE Trans. Power Syst., vol. 37, no. 2, pp. 1007–1017, 2022.
[23] H. Liu and W. Wu, "Online multi-agent reinforcement learning for decentralized inverter-based Volt-VAR control," IEEE Trans. Smart Grid, vol. 12, no. 4, pp. 2980–2990, 2021.
[24] D. Qiu, T. Chen, G. Strbac, and S. Bu, "Coordination for multienergy microgrids using multiagent reinforcement learning," IEEE Trans. Ind. Informat., vol. 19, no. 4, pp. 5689–5700, 2023.
[25] Y. Wu, T. Zhao, H. Yan, M. Liu, and N. Liu, "Hierarchical hybrid multi-agent deep reinforcement learning for peer-to-peer energy trading among multiple heterogeneous microgrids," IEEE Trans. Smart Grid, vol. 14, no. 6, pp. 4649–4665, 2023.
[26] T. Qian, C. Shao, X. Li, X. Wang, Z. Chen, and M. Shahidehpour, "Multi-agent deep reinforcement learning method for EV charging station game," IEEE Trans. Power Syst., vol. 37, no. 3, pp. 1682–1694, 2022.
[27] P. Chen, S. Liu, X. Wang, and I. Kamwa, "Physics-guided multi-agent adversarial reinforcement learning for robust active voltage control with peer-to-peer energy trading," IEEE Trans. Power Syst., vol. 39, no. 6, pp. 7089–7101, 2024.
[28] Y. Zhou, Z. Ma, T. Wang, J. Zhang, X. Shi, and S. Zou, "Joint energy and carbon trading for multi-microgrid system based on multi-agent deep reinforcement learning," IEEE Trans. Power Syst., vol. 39, no. 6, pp. 7376–7388, 2024.
[29] S. Li, D. Cao, W. Hu, Q. Huang, Z. Chen, and F. Blaabjerg, "Multi-energy management of interconnected multi-microgrid systems using multi-agent deep reinforcement learning," J. Mod. Power Syst. Clean Energy, vol. 11, no. 5, pp. 1606–1617, 2023.
[30] S. Gu, L. Yang, Y. Du, G. Chen, F. Walter, J. Wang, and A. Knoll, "A review of safe reinforcement learning: Methods, theory and applications," IEEE Trans. Pattern Anal. Mach. Intell., 2024.
[31] L. Brunke, M. Greeff, A. W. Hall, Z. Yuan, S. Zhou, J. Panerati, and A. P. Schoellig, "Safe learning in robotics: From learning-based control to safe reinforcement learning," Annu. Rev. Control Robot. Auton. Syst., vol. 5, pp. 411–444, 2022.
[32] H. Krasowski, J. Thumm, M. Müller, L. Schäfer, X. Wang, and M. Althoff, "Provably safe reinforcement learning: Conceptual analysis, survey, and benchmarking," Trans. Mach. Learn. Res., 2023.
[33] V.-H. Bui, S. Das, A. Hussain, G. V. Hollweg, and W. Su, "A critical review of safe reinforcement learning techniques in smart grid applications," arXiv:2409.16256, 2024.
[34] M. Alshiekh, R. Bloem, R. Ehlers, B. Könighofer, S. Niekum, and U. Topcu, "Safe reinforcement learning via shielding," in Proc. AAAI Conf. Artif. Intell., vol. 32, 2018.
[35] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," in Proc. Int. Conf. Mach. Learn. (ICML), 2017, pp. 22–31.
[36] E. Altman, Constrained Markov Decision Processes. Boca Raton, FL, USA: CRC Press, 1999.
[37] J. Zhou, M. Yang, Y. Zhan, and L. Xu, "Digital twin application for reinforcement learning based optimal scheduling and reliability management enhancement of systems," Solar Energy, vol. 252, pp. 29–38, 2023.
[38] Y. Ye, B. Xu, H. Wang, J. Zhang, B. Lawler, and B. Ayalew, "Deep reinforcement learning-based energy management system enhancement using digital twin for electric vehicles," Energy, vol. 312, art. 133384, 2024.
[39] M. Pan, Q. Xing, Z. Chai, H. Zhao, Q. Sun, and D. Duan, "Real-time digital twin machine learning-based cost minimization model for renewable-based microgrids considering uncertainty," Solar Energy, vol. 250, pp. 355–367, 2023.
[40] D. Qi, X. Xi, Y. Tang, Y. Zheng, and Z. Guo, "Real-time scheduling of power grid digital twin tasks in cloud via deep reinforcement learning," J. Cloud Comput., vol. 13, no. 1, art. 121, 2024.
[41] K. Xia, C. Sacco, M. Kirkpatrick, C. Saidy, L. Nguyen, A. Kircaliali, and R. Harik, "A digital twin to train deep reinforcement learning agent for smart manufacturing

plants: Environment, interfaces and intelligence," J. Manuf. Syst., vol. 58, pp. 210–230, 2021.
[42] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," arXiv:1707.06347, 2017.
[43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 5998–6008.