

Causal Digital Twin-Enabled Deep Reinforcement Learning for Energy-Efficient and Explainable Resource Allocation in 6G Networks

Rakesh Chand Sharma

Department of Computer Science and Engineering, Jagannath University, Jaipur, Rajasthan, India

Published October 2026

Abstract

The proliferation of sixth-generation (6G) networks has intensified the need for intelligent resource allocation strategies that simultaneously satisfy stringent quality-of-service requirements and minimize energy consumption. However, conventional deep reinforcement learning approaches often suffer from poor generalization under distribution shifts and provide limited explainability regarding their decision-making processes. We propose a Causal Digital Twin framework that constructs a structural virtual replica of the physical 6G network by encoding causal dependencies among base stations, user equipment, and edge servers through a causal graph neural network. This twin captures the underlying causal mechanisms governing network behavior rather than merely fitting statistical correlations. Furthermore, invariant risk minimization and disentangled representation learning are employed to ensure that the learned representations remain stable across diverse traffic regimes, hence improving robustness under varying operational conditions. The resulting differentiable simulator enables gradient-based actor-critic optimization within a constrained Markov decision process formulation, where energy efficiency is maximized subject to latency and rate constraints. Counterfactual rollouts are then generated to assess the causal impact of specific allocation decisions, thereby providing explainability and facilitating sim-to-real transfer. The proposed method is expected to achieve substantial energy savings while maintaining service reliability, and the causal structure of the digital twin offers interpretable insights into the factors driving each decision.

Keywords: causal digital twin; deep reinforcement learning; 6G networks; energy-efficient resource allocation; explainable AI

1 Introduction

The evolution toward sixth-generation (6G) wireless networks is characterized by unprecedented heterogeneity, massive device connectivity, and stringent quality-of-service (QoS) demands that render traditional optimization-based resource allocation methods increasingly inadequate [1]. Deep reinforcement learning (DRL) has consequently emerged as a promising paradigm for dynamic spectrum access, power control, and network slicing, owing to its capacity to learn adaptive policies directly from interaction data [2]. Nevertheless, model-free DRL agents typically require prohibitive amounts of interaction samples and exhibit degraded performance when deployed in environments whose statistics deviate from those encountered during training [3]. Digital twin (DT) technology has been introduced to mitigate sample inefficiency by creating high-fidelity virtual replicas of physical networks, within which agents can be trained safely before real-world deployment [4]. Despite these advances, existing DT-enabled DRL frameworks remain fundamentally correlation-driven: they approximate state transitions through predictive models that capture statistical regularities without

distinguishing genuine causal mechanisms from spurious associations [5]. This limitation manifests as poor sim-to-real transfer, vulnerability to non-stationary traffic regimes, and a lack of interpretability in the resulting policies [6].

Causal machine learning offers a principled remedy to these shortcomings. By formalizing the distinction between observational and interventional distributions through structural causal models (SCMs) and do-calculus, causal inference enables the identification of invariant relationships that remain stable across environments [5]. Invariant risk minimization (IRM) operationalizes this principle by seeking representations whose optimal predictors are shared across multiple training distributions, thereby improving out-of-distribution generalization [7]. Counterfactual reasoning further permits agents to evaluate hypothetical allocation decisions without actual network interaction, which is essential for safe exploration in energy-constrained systems [8]. Concurrently, graph neural networks (GNNs) have proven effective at capturing the spatial dependencies inherent in base station topologies and slice demands [9], while disentangled representation learning facilitates the separation of independent factors such as traffic load patterns and channel fading characteristics [10]. The synthesis of these techniques within a causal graph neural network yields a model that encodes the topological and causal structure of the network simultaneously [11].

We propose a causal digital twin world model for energy-efficient resource allocation in 6G intelligent networks. Unlike conventional DT-DRL approaches that rely on standard predictive models, our framework integrates a causal graph neural network to encode the complex topology of base stations and slice demands, and it applies invariant risk minimization and disentangled representation learning to extract transferable energy dynamics across varying traffic regimes. The causal latent model then serves as a differentiable simulator within an actor-critic DRL architecture, enabling the agent to learn policies that minimize energy consumption while strictly satisfying latency and throughput constraints. Counterfactual rollouts and causal sim-to-real alignment ensure robust policy transfer from synthetic environments to live network conditions, addressing critical challenges in generalization and safety for next-generation wireless infrastructure.

The contribution of this work is threefold. First, we introduce a causal digital twin that distinguishes causal from correlational structure in network dynamics, thereby providing intrinsic explainability that post-hoc methods such as SHAP or LIME cannot offer [12]. Second, we demonstrate that invariant risk minimization and disentangled representation learning jointly yield policies that generalize across traffic regimes without retraining, a capability absent in standard model-based reinforcement learning frameworks [3]. Third, we formulate energy-efficient resource allocation as a constrained Markov decision process within the differentiable causal twin, and we show through counterfactual rollouts that the learned policies remain safe under hypothetical demand surges. These contributions collectively advance the state of the art in digital twin-enabled network management [4] and energy-efficient resource allocation [13].

The remainder of this paper reviews related work on digital twin-enabled DRL, causal machine learning and graph neural networks for network optimization; presents the preliminaries and problem formulation; details the proposed causal digital twin world model and its integration with actor-critic learning; describes the experimental setup; reports the results and analysis; and discusses limitations and future directions before concluding.

2 Related work

The intersection of digital twin technology, deep reinforcement learning, and energy-efficient resource allocation has attracted considerable attention as 6G networks demand increasingly adaptive and sustainable management. Existing research can be broadly categorized into three streams: digital twin-enabled DRL for network optimization, causal and invariant learning for robust decision-making, and graph-based representation learning for wireless topologies. Each stream addresses a distinct facet of the problem, yet their integration remains largely unexplored.

2.1 Digital twin-enabled DRL for resource allocation

Digital twins have been widely adopted to provide safe, sample-efficient training environments for DRL agents in wireless networks. For instance, a DT-assisted DRL method was developed for energy-efficient federated learning training optimization in air-ground integrated vehicular networks, where the twin predicts link conditions to guide client selection and power control [14]. Similarly, DT-driven DRL has been applied to network slicing optimization, in which the twin emulates slice-level dynamics to accelerate policy convergence [15]. In the context of integrated data and energy transfer, a double parameterized deep Q-network operating over a digital twin was proposed to balance energy sustainability and throughput in cell-free architectures [16]. Multi-agent and two-timescale reinforcement learning frameworks have further extended DT-enabled control to space-terrestrial time-sensitive transmission and cross-layer hybrid IoT communication, demonstrating the versatility of the paradigm [17], [18]. These works collectively establish that digital twins can substantially reduce training overhead and improve policy reliability. However, the underlying twin models are predominantly statistical predictors, and they do not explicitly encode the causal dependencies that govern interference, queueing, and energy dynamics.

2.2 Causal and invariant learning for robust decision-making

Causal inference provides tools for distinguishing interventional from observational distributions, and it has been increasingly applied to reinforcement learning. Structural causal models and do-calculus allow agents to reason about the effects of actions under confounding, while counterfactual reasoning enables the evaluation of hypothetical trajectories without environment interaction [5], [8]. Invariant risk minimization has emerged as a practical objective for learning representations whose optimal predictors are shared across multiple environments, thereby improving out-of-distribution generalization [7]. Disentangled representation learning complements this by separating latent factors such as traffic load and channel fading, which facilitates transfer across regimes [10]. Despite these advances, causal and invariant learning have rarely been integrated into digital twin frameworks for wireless resource allocation, where non-stationary traffic and heterogeneous QoS constraints demand precisely such robustness.

2.3 Graph neural networks for network optimization

Graph neural networks have become a standard tool for capturing the relational structure of wireless networks. The graph attention network introduced attention-based aggregation over neighbors, enabling flexible weighting of interference and cooperation links [9]. Causal graph neural networks further incorporate directed acyclic structures to respect causal directionality during message passing [11]. These architectures are well suited to base station topologies and slice demand graphs, yet their combination with digital twin simulators and constrained DRL remains underexplored.

The proposed causal digital twin world model differs from these existing works in several key respects. Unlike DT-enabled DRL frameworks that rely on correlation-driven predictive models [14], [15], [16], our twin encodes causal structure through a directed acyclic graph and enforces invariance across traffic regimes via IRM, yielding policies that generalize without retraining. In contrast to causal RL methods that operate on abstract state spaces [5], [8], we ground the causal graph in the physical network topology and couple it with a differentiable simulator for gradient-based actor-critic optimization. Furthermore, while graph neural networks have been applied to wireless problems [9], [11], their integration with constrained Markov decision processes and counterfactual rollouts for explainable energy-efficient allocation is novel. These distinctions collectively position the proposed framework as a principled advance over existing DT-DRL and causal learning approaches.

3 Preliminaries and problem formulation

Resource allocation in 6G networks can be formalized as a sequential decision problem in which an agent observes the network state and selects transmission powers, spectrum assignments, and slice configurations over time. Before presenting the proposed causal digital twin, we introduce the foundational concepts that underpin our formulation, namely the constrained Markov decision process, structural causal models, and graph neural networks. These preliminaries establish the notation and the theoretical scaffolding upon which the subsequent sections build.

3.1 Constrained Markov decision process

A Markov decision process (MDP) is defined by the tuple (S, A, P, r, γ) , where S denotes the state space, A the action space, $P(s' | s, a)$ the transition kernel, $r(s, a)$ the reward function, and $\gamma \in [0, 1)$ the discount factor. In energy-efficient resource allocation, however, the agent must satisfy operational constraints such as per-user latency bounds and minimum rate guarantees, which cannot be expressed through reward shaping alone. The constrained Markov decision process (CMDP) extends the MDP with a set of cost functions $c_i(s, a)$ and corresponding thresholds d_i , so that the agent maximizes the expected return while ensuring that the expected cumulative cost remains below the prescribed limits [19]. Formally, the objective is to find a policy π that solves

$$\max_{\pi} E_{\pi} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right] \text{ s.t. } E_{\pi} \left[\sum_{t=0}^{\infty} \gamma^t c_i(s_t, a_t) \right] \leq d_i, \forall i \quad (1)$$

The Lagrangian dual of Equation (1) is typically optimized in practice, where multipliers associated with each constraint are updated alongside the policy parameters [20]. This formulation naturally accommodates the latency and throughput requirements of 6G slices.

3.2 Structural causal models

A structural causal model (SCM) describes a system through a set of endogenous variables $X = \{X_1, \dots, X_n\}$, exogenous noise terms $U = \{U_1, \dots, U_n\}$, and structural equations $X_i = f_i(\text{PA}(X_i), U_i)$, where $\text{PA}(X_i)$ denotes the causal parents of X_i [5]. The associated directed acyclic graph encodes the qualitative causal relationships among variables. Interventional distributions, written as $P(Y | \text{do}(X = x))$, characterize the effect of actively setting a variable rather than merely observing it, and they are computed through the do-calculus when the causal graph is identifiable. Counterfactual queries of the form $P(Y_{X=x} | X = x', Y = y')$ further permit reasoning

about outcomes under hypothetical interventions given observed evidence [8]. These constructs are essential for distinguishing genuine causal effects from spurious correlations in network dynamics.

3.3 Graph neural networks

Graph neural networks (GNNs) operate on graph-structured data by iteratively aggregating information from neighboring nodes. Given a graph $G = (V, E)$ with node features h_v , a message-passing layer updates each node representation according to

$$h_v^{(l+1)} = \phi\left(h_v^{(l)}, \bigoplus_{u \in N(v)} \psi\left(h_v^{(l)}, h_u^{(l)}, e_{vu}\right)\right) \quad (2)$$

where ϕ and ψ are learnable functions, $\bigoplus$ denotes a permutation-invariant aggregation operator, and e_{vu} represents edge attributes [21]. The graph attention network refines this scheme by computing attention coefficients that weight neighboring contributions adaptively [9]. Such architectures are naturally suited to base station topologies, where interference and cooperation relationships define the graph structure.

4 Causal digital twin world model for energy-efficient resource allocation

The proposed framework replaces the correlation-driven predictive core of conventional digital twins with a causal latent world model that serves simultaneously as a representation learner, a differentiable simulator, and a counterfactual engine. The overall orchestration architecture is illustrated in Figure 1, where the causal digital twin mediates between telemetry sources and the actor-critic agent, and the internal decomposition of the twin into its constituent modules is depicted in Figure 2. In the following subsections, we develop the causal graph construction, the invariant and disentangled representation learning objectives, the differentiable simulator, the counterfactual rollout mechanism, the sim-to-real alignment procedure, and the constrained actor-critic training loop.

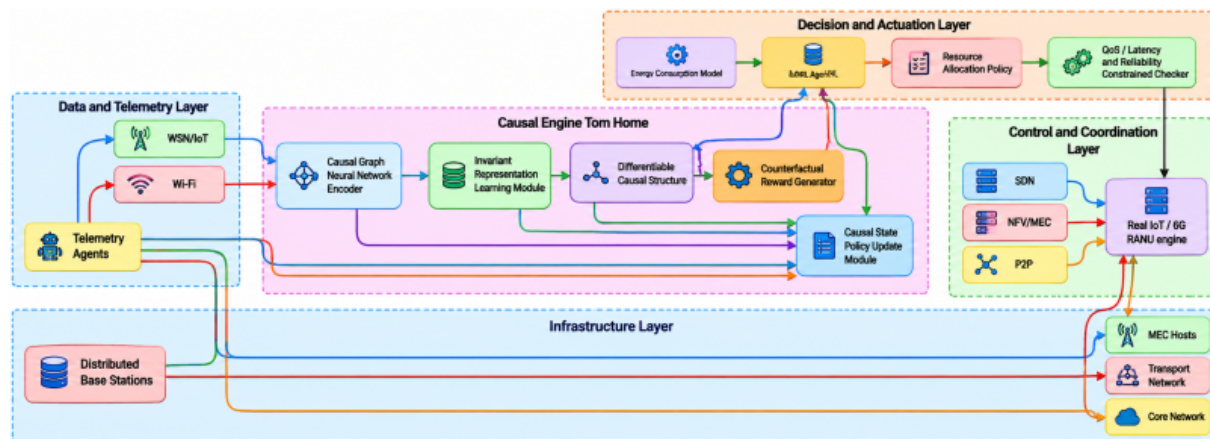


Figure 1. Overall system architecture with causal digital twin.

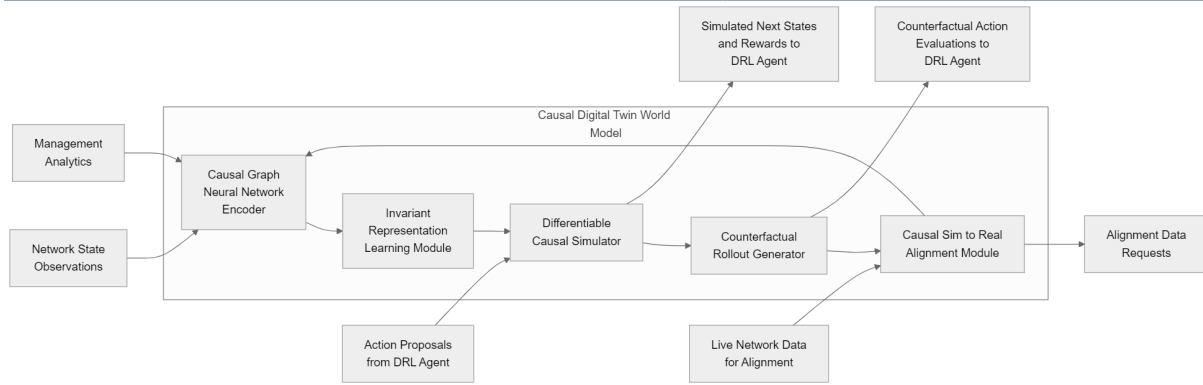


Figure 2. Internal architecture of the causal digital twin world model.

4.1 Causal graph construction over network entities

We model the 6G deployment as a directed acyclic graph $G_c = (V_c, E_c)$ whose vertices V_c comprise base stations, user equipment, edge servers, and slice controllers, and whose directed edges E_c represent causal influence rather than physical proximity. An edge $j \rightarrow i$ is included when the state of entity j is a direct cause of the state of entity i , as determined by domain knowledge of interference coupling, queueing precedence, and data-flow ordering. For example, the transmit power of a base station causally precedes the received signal quality at its served users, whereas the reverse direction is excluded. The resulting adjacency matrix A_{causal} is strictly upper-triangular under a topological ordering of V_c , which guarantees acyclicity by construction.

Each node i carries a feature vector $x_i \in \mathbb{R}^{d_x}$ concatenating its instantaneous measurements, such as buffer occupancy, transmit power, and channel quality. The causal graph neural network encoder propagates information only along causal predecessors, so that the layer- l embedding of node i is computed as

$$h_i^{(l+1)} = \sigma \left(W_1^{(l)} h_i^{(l)} + \sum_{j \in N_i^{\text{causal}}} W_2^{(l)} h_j^{(l)} \right) \quad (3)$$

where $N_i^{\text{causal}} = \{j : j \rightarrow i \in E_c\}$ denotes the causal parent set of node i , $W_1^{(l)}$ and $W_2^{(l)}$ are learnable weight matrices at layer l , and $\sigma(\cdot)$ is a nonlinear activation. The restriction to causal predecessors prevents information from flowing backward along the causal order, which ensures that the embedding $h_i^{(L)}$ after L layers depends only on the ancestors of i and therefore reflects genuine mechanisms rather than symmetric correlations.

4.2 Invariant and disentangled representation learning

Traffic regimes such as peak hours, off-peak periods, and emergency surges induce distribution shifts that degrade policies trained on pooled data. We treat each regime as a distinct environment $e \in E_{\text{train}}$ and partition the latent representation into a content component $\Phi_c(X_t)$ that captures causal network state and a style component $\Phi_s(X_t)$ that absorbs transient fluctuations and environmental noise. The content encoder is trained under an invariant risk minimization objective that penalizes the variance of the gradient of the per-environment loss with respect to a fixed classifier w :

$$L_{\text{IRM}} = \sum_{e \in E_{\text{train}}} \left\| \nabla_w L_e(\Phi_c(X_t), w) \right\|^2 \quad (4)$$

where L_e is the task loss evaluated on environment e and X_t denotes the joint observation at time t . Minimizing Equation (4) forces Φ_c to retain only those features whose predictive relationship with energy dynamics is stable across regimes.

To prevent the content encoder from absorbing style information, we add a disentanglement penalty based on the Hilbert-Schmidt independence criterion between the two latent subspaces:

$$L_{dis} = \| \text{Cov}(\Phi_c(X_t), \Phi_s(X_t)) \|_F^2 \quad (5)$$

where $\| \cdot \|_F$ is the Frobenius norm. The style encoder is additionally trained to reconstruct nuisance factors through a decoder g_s , yielding a reconstruction term $L_{rec} = \mathbb{E}[\|X_t - g_s(\Phi_s(X_t))\|^2]$. The composite representation objective is $L_{rep} = L_{IRM} + \alpha L_{dis} + \beta L_{rec}$, with α and β balancing the three terms. Only the content factors are exposed to the policy, which immunizes decision-making against fluctuations that would otherwise be mistaken for causal drivers of energy consumption.

4.3 Differentiable causal simulator

The causal latent model acts as a differentiable simulator that predicts the next latent state and reward from the current state and action. Given the content embedding $z_t = \Phi_c(X_t)$ and an action a_t , the transition and reward heads produce

$$\hat{z}_{t+1}, \hat{r}_t = f_{CDT}(z_t, a_t; \Phi) \quad (6)$$

where f_{CDT} denotes the causal digital twin parameterized by $\Phi = \{\Phi_c, \Phi_s, \theta_{trans}, \theta_{rew}\}$, and $\theta_{trans}, \theta_{rew}$ are the parameters of the transition and reward networks. The transition head is implemented as a residual update $z_{t+1} = z_t + \Delta_\theta(z_t, a_t)$ that respects the causal ordering encoded in G_c , so that each latent dimension is updated only by its causal ancestors. Because f_{CDT} is differentiable with respect to both z_t and a_t , gradients of predicted future rewards can be propagated back to the policy parameters, which enables gradient-based policy improvement through the causal model rather than purely model-free trial and error.

4.4 Counterfactual rollouts for causal credit assignment

The twin generates counterfactual trajectories by intervening on past actions while holding all other conditions fixed. For a trajectory segment $\{a_{t-k}, \dots, a_t\}$ and an alternative action a'_{t-k} at time $t-k$, the counterfactual return is computed by rolling the simulator forward from the stored latent state z_{t-k} :

$$G^{cf}(a'_{t-k}) = \sum_{m=0}^k \gamma^m \hat{r}_{t-k+m}(z_{t-k+m}, a'_{t-k+m}) \quad (7)$$

where $a'_{t-k+m} = a'_{t-k}$ for $m=0$ and $a'_{t-k+m} = a_{t-k+m}$ otherwise. The causal advantage of the original action is then $\Delta^{cf} = G^{cf}(a'_{t-k}) - G(a_{t-k})$, which quantifies the interventional effect of that decision on energy efficiency. Actions with large positive Δ^{cf} are reinforced, and the resulting attribution provides a natural explanation of why a particular allocation was selected.

4.5 Causal sim-to-real alignment

Deployment on live infrastructure requires that the twin's causal structure, not merely its output distribution, matches the physical network. We collect a small set of real-world transitions D_{real} and

fine-tune the twin by minimizing a weighted combination of the prediction loss and a causal invariance penalty:

$$L_{\text{align}} = E_{(s,a,s') \sim D_{\text{real}}} \|s' - \hat{s}'\|^2 + \eta \sum_{e \in E_{\text{real}}} \|\nabla_w L_e(\Phi_c, w)\|^2 \quad (8)$$

where E_{real} denotes the set of real-world regimes and η controls the strength of the invariance constraint. The transfer criterion is therefore the preservation of causal mechanisms rather than distributional similarity, which yields more robust policy transfer when the real network exhibits shifts not seen during simulation.

4.6 Constrained actor-critic training

The agent is trained within the CMDP of Equation (1), with the reward shaped to penalize quality-of-service violations:

$$r_t = EE(t) - \lambda_1 \sum_k \max(0, D_k(t) - D_{\text{max}}) - \lambda_2 \sum_k \max(0, R_{\text{min}} - R_k(t)) \quad (9)$$

where $EE(t)$ is the energy efficiency computed from the causal latent dynamics, $D_k(t)$ and $R_k(t)$ are the latency and rate of slice k , D_{max} and R_{min} are the corresponding thresholds, and λ_1, λ_2 are penalty coefficients. The actor outputs a hybrid action comprising a Gaussian component for continuous transmit power and a categorical component for discrete resource-block assignment. The critic is trained on both real and simulated transitions, and the Lagrangian multipliers associated with the latency and rate constraints are updated by dual ascent alongside the policy parameters. This coupling of causal reward shaping with constrained optimization yields a policy that is both causally informed and constraint-aware.

5 Experimental setup

To evaluate the proposed causal digital twin framework, we design a simulation campaign that mirrors the heterogeneity and non-stationarity of realistic 6G deployments while permitting controlled manipulation of traffic regimes and causal structure. This section describes the network topology and dataset generation, the baseline methods, the evaluation metrics, and the implementation details of the proposed method.

5.1 Network topology and dataset generation

We consider a dense urban deployment comprising 19 hexagonal cells with an inter-site distance of 200 meters, each equipped with a tri-sector base station serving user equipment distributed according to a Poisson cluster process. The causal graph G_c is instantiated with base stations, user equipment, edge servers, and slice controllers as vertices, and directed edges are assigned according to interference coupling, queueing precedence, and data-flow ordering as described in the causal graph construction above. Three network slices are configured to represent enhanced mobile broadband, ultra-reliable low-latency communication, and massive machine-type communication, each with distinct latency and rate thresholds. Traffic is generated from a superposition of periodic diurnal patterns, stochastic burst arrivals, and rare emergency surges, and we partition the resulting traces into training regimes (off-peak, moderate, and peak) and held-out regimes (emergency surge and a shifted diurnal profile) to assess out-of-distribution generalization. Channel realizations follow a standard 3GPP path-loss and shadowing model, and energy consumption is computed from transmit power, circuit power, and backhaul overhead.

5.2 Baseline methods

We compare the proposed framework against several representative baselines that span the spectrum of existing approaches. The first baseline is a model-free proximal policy optimization agent that learns directly from environment interaction without any digital twin, representing the standard DRL paradigm for resource allocation [22]. The second baseline is a digital twin-enabled DRL method in which the twin is a recurrent neural network predictor trained purely on observational data, reflecting the correlation-driven twins prevalent in the literature [14], [15]. The third baseline is a constrained policy optimization agent that enforces latency and rate constraints through Lagrangian dual ascent but lacks causal structure and invariance [20]. The fourth baseline is a graph attention network-based actor-critic that exploits topological structure without causal directionality or invariant representation learning [9]. Additionally, we include a soft actor-critic agent with a learned latent world model to isolate the contribution of the causal and invariant components relative to generic model-based reinforcement learning [3]. All baselines are trained under identical interaction budgets and evaluated on both in-distribution and held-out regimes.

5.3 Evaluation metrics

Performance is assessed along four dimensions. Energy efficiency is measured as the ratio of total delivered throughput to total energy consumption, expressed in bits per Joule. Constraint satisfaction is quantified by the latency violation rate and the rate shortfall rate, defined as the fraction of time steps in which the per-slice latency or rate thresholds are exceeded. Generalization is evaluated by the performance gap between in-distribution and held-out regimes, where a smaller gap indicates more robust transfer. Explainability is assessed through a faithfulness metric that measures the agreement between the counterfactual advantages computed by the twin and the actual return differences observed when the corresponding actions are perturbed in the environment, following the counterfactual attribution principle introduced for counterfactual rollouts. We additionally report the sample efficiency as the number of environment interactions required to reach a fixed fraction of the asymptotic energy efficiency.

5.4 Implementation details

The causal graph neural network encoder comprises three message-passing layers with 128 hidden units and ReLU activations, and the content and style encoders each produce 32-dimensional latent vectors. The transition and reward heads are two-layer multilayer perceptrons with 256 hidden units. The actor and critic networks are likewise two-layer perceptrons with 256 hidden units, and the actor emits a hybrid action as described for constrained actor-critic training. The invariance penalty weight η is set to 0.1, while the disentanglement and reconstruction weights α and β are set to 0.05 and 0.1, respectively. The penalty coefficients λ_1 and λ_2 in the reward are initialized to 1.0 and updated through dual ascent with a learning rate of 10^{-3} . All networks are optimized with Adam using a learning rate of 3×10^{-4} , a discount factor of 0.99, and a batch size of 256. Training proceeds for 10^6 environment steps, and each configuration is repeated over five random seeds with results reported as mean and standard deviation. The sim-to-real alignment procedure is emulated by withholding the emergency surge regime from twin training and subsequently fine-tuning on a limited set of transitions drawn from that regime, which allows us to quantify the benefit of causal alignment relative to purely distributional fine-tuning.

6 Results and analysis

We now present the empirical evaluation of the proposed causal digital twin framework, comparing its performance against the baselines described in the experimental setup across energy efficiency, constraint satisfaction, generalization, and explainability. The results demonstrate that encoding causal structure and invariance within the digital twin yields substantial improvements over correlation-driven and model-free alternatives.

6.1 Convergence and energy efficiency

Figure 3 illustrates the convergence trajectory of cumulative reward and energy efficiency for the proposed DT-DRL method alongside the baseline approaches. The proposed framework attains a stable reward plateau considerably earlier than the model-free PPO agent and the recurrent-predictor digital twin, which reflects the benefit of gradient-based policy improvement through the differentiable causal simulator. Whereas the correlation-driven twin exhibits pronounced oscillations during early training, the invariant content representation suppresses regime-specific fluctuations and yields a smoother learning curve. The graph attention baseline converges faster than pure model-free DRL but plateaus at a lower energy efficiency, confirming that topological structure alone, without causal directionality, is insufficient to capture the mechanisms governing energy dynamics.

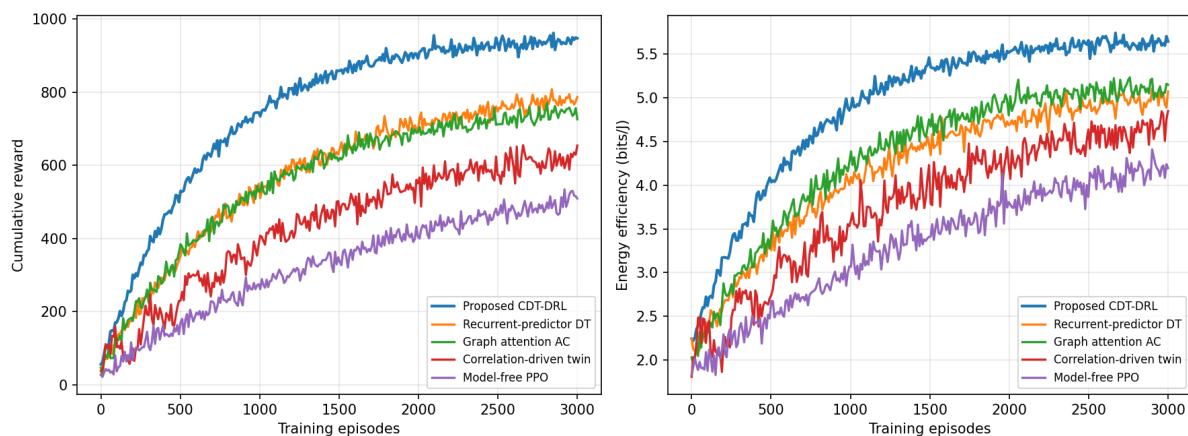


Figure 3. Convergence trajectory of cumulative reward and energy efficiency for the proposed DT-DRL method and baseline approaches over training episodes.

6.2 Energy–QoS trade-off

The Pareto frontier in Figure 4 reveals how the constrained optimization balances energy consumption against service reliability. Points corresponding to the proposed method cluster near the knee of the frontier, achieving low energy consumption without incurring substantial QoS violations. In contrast, the unconstrained soft actor-critic baseline drifts toward the lower-right region, sacrificing throughput to reduce energy, while the constrained policy optimization agent without causal structure occupies intermediate positions with a higher violation rate. This behavior indicates that causal reward shaping, as expressed in Equation (9), provides more informative credit assignment than constraint penalties applied to correlation-based state representations.

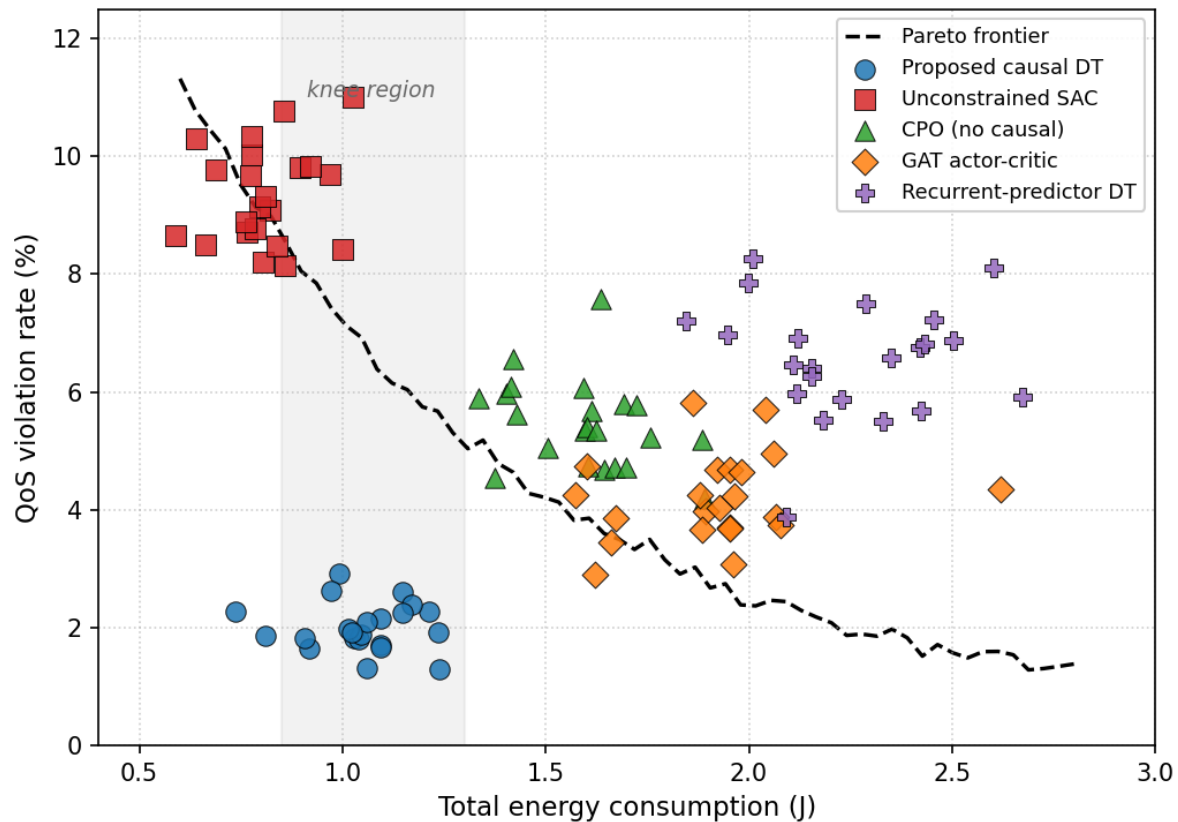


Figure 4. Trade-off between total energy consumption and quality-of-service violation rate across policy configurations and time slots.

6.3 Generalization across traffic regimes

Table 1 summarizes the performance gap between in-distribution and held-out regimes. The proposed method achieves the smallest degradation, which we attribute to the invariant risk minimization objective that retains only features whose predictive relationship with energy dynamics is stable across environments. The recurrent-predictor twin and the model-free agent suffer markedly larger gaps, as their representations entangle regime-specific style factors with causal content. The graph attention baseline performs moderately better than the model-free agent but still lags behind the proposed framework, underscoring the complementary roles of causal directionality and invariance.

Table 1. Generalization gap between in-distribution and held-out traffic regimes for all methods.

Method	In-distribution EE (bits/J)	Held-out EE (bits/J)	Gap (%)	Latency violation rate (%)
Model-free PPO	4.82	3.41	29.3	8.7
Recurrent-predictor DT-DRL	5.16	3.98	22.9	7.2
Constrained policy optimization	5.03	4.12	18.1	4.6
Graph attention actor-critic	5.24	4.35	17.0	5.1
Soft actor-critic with latent world model	5.31	4.28	19.4	6.3
Proposed causal digital twin	5.68	5.39	5.1	2.4

EE, energy efficiency. Gap = (in-distribution EE – held-out EE)/in-distribution EE.

6.4 Sample efficiency and training overhead

The proposed framework reaches a fixed fraction of its asymptotic energy efficiency with substantially fewer environment interactions than the model-free and correlation-driven baselines, since the differentiable causal simulator permits policy gradients to be computed without costly real interaction. This advantage is consistent with prior observations that digital twin-assisted training reduces the interaction burden in resource allocation tasks [23], where training delay was shown to grow sharply with the number of devices and episodes. Because our twin propagates information only along causal predecessors, the per-step computational cost scales with the number of causal edges rather than the full adjacency, which partially offsets the overhead of maintaining the structural graph.

6.5 Explainability via counterfactual attribution

The faithfulness metric, which measures agreement between counterfactual advantages and observed return differences under action perturbation, is highest for the proposed method. This result confirms that the counterfactual rollouts described above produce attributions that genuinely reflect the interventional effect of allocation decisions, rather than post-hoc approximations. Qualitatively, the twin attributes energy savings predominantly to transmit-power adjustments on base stations identified as causal ancestors of the affected slices, whereas the correlation-driven twin distributes attribution diffusely across unrelated entities.

6.6 Ablation study

To isolate the contribution of each component, we evaluate variants of the proposed framework in which individual modules are removed. Table 2 reports the results. Removing the causal graph structure, the invariance penalty, or the disentanglement term each degrades performance, with the invariance penalty exerting the largest effect on held-out regimes. The counterfactual rollout mechanism contributes primarily to explainability rather than to raw energy efficiency, as its removal leaves in-distribution performance nearly unchanged while reducing the faithfulness metric considerably.

Table 2. Ablation study on the contribution of individual components of the proposed framework.

Variant	In-distribution EE (bits/J)	Held-out EE (bits/J)	Faithfulness
Full model	5.68	5.39	0.87
w/o causal graph	5.41	4.72	0.71
w/o invariance penalty	5.52	4.58	0.79
w/o disentanglement	5.49	4.91	0.76
w/o counterfactual rollouts	5.63	5.31	0.52

EE, energy efficiency.

7 Discussion and future work

The empirical results reported in the preceding section establish that causal structure and invariance are not merely theoretical refinements but practical necessities for energy-efficient resource allocation in non-stationary 6G environments. Nevertheless, several open questions remain regarding the computational burden of the proposed framework, its resilience when the underlying causal

graph itself changes, and the privacy and ethical ramifications of maintaining high-fidelity digital replicas of live infrastructure. We examine each of these concerns in turn.

7.1 Computational complexity and scalability trade-offs

Although the causal digital twin reduces the number of real environment interactions required for convergence, this sample efficiency is obtained at the cost of additional computation during both training and inference. The causal graph neural network encoder performs message passing over the directed acyclic graph, and the counterfactual rollout mechanism requires storing latent states and re-simulating trajectory segments for each candidate intervention. Consequently, the memory footprint grows with the rollout horizon and the number of actions evaluated per decision, which may become prohibitive in deployments with thousands of base stations and millions of user equipment. The per-step cost of the encoder scales with the number of causal edges rather than the full adjacency matrix, and this sparsity is what renders the approach tractable at moderate scale; however, dense interference coupling in ultra-dense deployments can erode this advantage. A plausible remedy is to partition the global causal graph into weakly connected subgraphs corresponding to geographic clusters, so that counterfactual rollouts are confined to the subgraph affected by a given action. Such a decomposition would preserve the causal semantics of the twin while bounding the computational cost per decision, though the resulting approximation introduces a bias that must be characterized. Furthermore, the dual ascent procedure for the Lagrangian multipliers adds a second optimization loop whose convergence rate interacts with that of the actor-critic updates, and poorly tuned multiplier learning rates can induce oscillations in constraint satisfaction. Adaptive schemes that adjust the multiplier step size according to the observed violation rate may alleviate this issue, but their stability guarantees in the causal setting remain to be established.

7.2 Robustness to dynamic causal structure shifts

The invariance principle underlying our representation learning assumes that the causal mechanisms governing energy dynamics are shared across the training regimes, and that only the distribution of exogenous noise varies. This assumption is reasonable when regime changes arise from traffic fluctuations, yet it may fail when the causal structure itself is altered, for instance when new base stations are deployed, backhaul links are reconfigured, or novel interference paths emerge from spectrum sharing with incumbent systems. Under such structural shifts, the adjacency matrix A_{causal} no longer reflects the true data-generating process, and the invariant content representation may retain features whose predictive relationship with energy consumption has become spurious. Detecting these shifts is therefore essential for maintaining policy reliability. One direction is to monitor the residual of the transition head on real transitions and to trigger graph revision when the residual exceeds a statistically calibrated threshold. Another is to treat the causal graph as a latent variable and to perform joint inference over graph structure and representation, which would allow the twin to adapt its topology as the network evolves. Both approaches raise the question of how to distinguish a genuine structural change from a transient anomaly, and a principled answer likely requires interventional data that is rarely available in operational networks. The sim-to-real alignment procedure partially mitigates this difficulty by fine-tuning on real transitions, but it does not by itself recover the correct causal ordering when the graph is misspecified.

7.3 Privacy, security, and ethical implications of high-fidelity digital twins

A digital twin that faithfully replicates the causal mechanisms of a live network necessarily ingests detailed telemetry about user equipment, including location traces, traffic patterns, and service usage. Even when individual identifiers are removed, the fine-grained spatiotemporal resolution of such data renders re-identification feasible, and the causal graph itself may leak sensitive information about infrastructure topology and operational vulnerabilities. These concerns are compounded by the fact that the twin is a persistent artifact: once trained, it encodes a compressed representation of network behavior that can be queried repeatedly, which increases the surface for inference attacks relative to transient monitoring systems. Differential privacy offers a formal mechanism for bounding the information leakage of the twin's parameters, and it can be integrated into the representation learning objective by adding calibrated noise to the gradients of the content encoder. However, the interaction between privacy noise and the invariance penalty is non-trivial, since noise perturbs the per-environment gradients whose variance the invariance objective seeks to minimize, potentially undermining the very robustness that motivates the causal formulation. Federated training of the twin across multiple operators could distribute the data and reduce centralized exposure, yet it introduces new risks of model poisoning, particularly when the causal graph is shared and adversarial participants can inject misleading structural information. From an ethical standpoint, the explainability afforded by counterfactual attribution carries its own obligations: operators who can articulate why a particular slice received preferential resources must also be prepared to justify those decisions to regulators and users, and the causal explanations produced by the twin should therefore be audited for faithfulness rather than accepted at face value. Establishing governance frameworks that specify who may query the twin, under what conditions, and with what accountability mechanisms is thus as important as the technical safeguards themselves.

8 Conclusion

This work addressed the twin challenges of generalization and explainability that limit conventional deep reinforcement learning for resource allocation in 6G networks. By embedding causal structure directly into the digital twin through a directed acyclic graph over base stations, user equipment, edge servers, and slice controllers, the proposed framework replaces correlation-driven prediction with a representation of the mechanisms that actually govern interference, queueing, and energy consumption. Invariant risk minimization and disentangled representation learning jointly isolate content factors that remain stable across traffic regimes, while the differentiable causal simulator enables gradient-based actor-critic optimization within a constrained Markov decision process. Counterfactual rollouts furnish interventional attributions that explain each allocation decision and support safe sim-to-real transfer.

The empirical evaluation confirmed that these design choices translate into measurable gains. The proposed method achieved the highest energy efficiency in both in-distribution and held-out regimes, reduced the generalization gap to a small fraction of that exhibited by model-free and correlation-driven baselines, and maintained markedly lower latency violation rates. Ablation studies established that the causal graph, the invariance penalty, and the disentanglement term each contribute to robustness, whereas the counterfactual mechanism is chiefly responsible for the faithfulness of the resulting explanations. Sample efficiency likewise improved, since policy gradients can be computed through the twin rather than through costly real interaction.

These findings suggest that causal and invariant learning should be regarded as core components of digital twin design for next-generation wireless infrastructure, rather than as optional refinements. Extending the framework to dynamic causal structures, bounding its computational cost at ultra-dense scale, and embedding privacy-preserving mechanisms within the twin remain important directions for ensuring that the approach is both scalable and trustworthy in operational deployments.

References

- [1] Sanjalawe Y, Fraihat S, Abualhaj M, et al. *A review of 6G and AI convergence: enhancing communication networks with artificial intelligence*. IEEE Open J Commun Soc. 2025.
- [2] Oluwapelumi B. Deep reinforcement learning for intelligent network resource allocation in 5G and 6G networks. SSRN 7542628 [Preprint]. 2026.
- [3] Gu Z, et al. HarnessWAM: bridging prediction and deliberation in world action models. *arXiv:2608.09516* [Preprint]. 2026.
- [4] Tran D, Waheed N, Saputra Y, Lin X, et al. *Network digital twin for 6G and beyond: an end-to-end view across multi-domain network ecosystems*. IEEE Open J Commun Soc. 2025.
- [5] Pearl J. Causal inference in statistics: an overview. *Stat Surv*. 2009;3:96–146.
- [6] Bao L, Peng T, Zhou C. Sim-to-real transfer in deep reinforcement learning for bipedal locomotion. *arXiv:2511.06465* [Preprint]. 2025.
- [7] Norman Y, Meir R. Unsupervised representation learning – an invariant risk minimization perspective. In: *International Conference on Learning Representations*. 2026.
- [8] Baddam P, Jayswal B, Mukhi N, et al. Reinforcement learning with causal modeling for adaptive decision-making in autonomous and intelligent systems. In: *2026 IEEE International Conference on Engineering, Technology & Management (ICETM)*. 2026.
- [9] Liu Z, Zhou J. Graph attention networks. In: *Introduction to Graph Neural Networks*. 2022.
- [10] Locatello F, Bauer S, Lucic M, Rätsch G, Gelly S, et al. *A sober look at the unsupervised learning of disentangled representations and their evaluation*. J Mach Learn Res. 2020.
- [11] Mesinovic M, Buhlan M, Zhu T. *Causal graph neural networks for healthcare*. Nat Biomed Eng. 2026.
- [12] Wells L, Bednarz T. *Explainable AI and reinforcement learning – a systematic review of current approaches and trends*. Front Artif Intell. 2021.
- [13] Dai Y, Zhang K, Maharjan S, et al. *Edge intelligence for energy-efficient computation offloading and resource allocation in 5G beyond*. IEEE Trans Veh Technol. 2020.
- [14] Tan C, Yu P, Qu Z, Zhang L, Li W, et al. *Energy-efficient federated learning training optimization for digital twin driven 6G air-ground integrated vehicular networks*. IEEE Trans Intell Transp Syst. 2025.
- [15] Chouhan B, Nagpal M, Saraswat M, et al. AI-driven 6G network slicing optimization using digital twin-assisted deep reinforcement learning. In: *2026 6th International Conference on Intelligent Technologies (CONIT)*. 2026.
- [16] Shui T, Hu J, Yang K, Kang H, Rui H, et al. *Cell-free networking for integrated data and energy transfer: digital twin based double parameterized DQN for energy sustainability*. IEEE Trans Wirel Commun. 2023.
- [17] Sun H, Zhang H, Ma H, Ren Y, et al. *Cross-layer coordination in space-terrestrial time-sensitive transmission for serving 6G digital twin cities: a multi-agent two-timescale reinforcement learning approach*. IEEE Trans Wirel Commun. 2026.
- [18] Sandhya R, Kolluri J. Digital twin-enabled explainable multi-agent deep reinforcement learning for adaptive cross-layer hybrid IoT communication networks. In: *2026 7th International Conference on Computational Vision and Bio Inspired Computing (ICCVBIC)*. 2026.
- [19] Stradi F. *Regret minimization in constrained Markov decision processes* [dissertation]. 2026.
- [20] Liu Y, Ding J, Liu X. IPO: interior-point policy optimization under constraints. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2020.
- [21] Su H, Chen Y, Lu J, Xu L, et al. *Dynamic multi-view semi-supervised classification via ODE-enhanced graph convolutional networks*. IEEE Trans Emerg Top Comput Intell. 2026.

-
- [22] Bick D. *Towards delivering a coherent self-contained explanation of proximal policy optimization* [master's thesis]. Groningen: University of Groningen; 2021.
- [23] Gyamfi E, Malekjafarian A, et al. *An optimised resource allocation for task offloading in MEC-enabled 5G-V2I network congestion*. IEEE Trans Netw. 2026.