Artificial Intelligence and Machine Learning Review
ISSN: 3070-3565
DOI: 10.69987/AIMLR.2025.60402

CarbonShift: Harnessing Grid Carbon Variability for Geo-Distributed Workload Scheduling

Yanhuan Chen[1], Zijie Chen[1.2], Danbing Zou[2]
[1] *Master of Engineering, Dartmouth College, NH, USA*
[1.2] *Computer Engineering, University of Toronto Master, Toronto, Canada*
[2] *Computer Science and Technology, Wuhan University, Wuhan, China*

Keywords

carbon-aware computing, geo-distributed scheduling, renewable energy, deep reinforcement learning

Abstract

Data centers account for 1-1.3% of total U.S. electricity consumption, with carbon emissions escalating rapidly due to artificial intelligence training and cloud computing demands. Current scheduling approaches prioritize performance and cost optimization while largely overlooking the substantial spatio-temporal variability in grid carbon intensity, which can differ by 5-10x across regions and time periods. This paper presents CarbonShift, a carbon-aware scheduling framework that exploits grid carbon intensity variations for geo-distributed workload management. The framework integrates workload energy profiling, LSTM-based carbon intensity forecasting, and deep reinforcement learning-driven optimization to balance carbon reduction with job completion time and data transfer costs. Experimental results demonstrate carbon emission reductions of 42-67% compared to carbon-agnostic scheduling while maintaining acceptable performance trade-offs.

1. Introduction

1.1 Background and Motivation

1.1.1 Rising energy consumption of U.S. data centers

Data centers currently consume between 1% and 1.3% of U.S. total electricity supply, with projections indicating potential growth to 3-9% by 2030 as artificial intelligence workloads and cloud adoption accelerate[1]. A single training run of large language models such as GPT-3 requires approximately 1,287 MWh, equivalent to 120 years of average American household electricity consumption. The carbon footprint of AI inference operations has become particularly concerning, with individual AI-powered search queries consuming 10 times the energy of traditional web searches. The concentration of computational workloads in specific geographic regions further exacerbates this challenge, as these locations may rely heavily on carbon-intensive power sources during peak demand periods.

1.1.2 Carbon emission challenges from AI and cloud computing growth

The environmental impact of computational workloads extends beyond operational energy consumption to encompass the full lifecycle of digital infrastructure. Recent analyses indicate that embodied carbon from server manufacturing and transportation accounts for 20-50% of a data center's total carbon footprint[2]. Cloud service providers operate geographically distributed data centers to ensure reliability, reduce latency, and optimize costs. This distributed architecture creates opportunities for carbon-aware workload placement, yet current scheduling algorithms predominantly focus on resource utilization efficiency and user-perceived performance. Grid carbon intensity fluctuates substantially throughout the day as renewable energy generation varies with solar irradiance and wind patterns.

1.1.3 Policy drivers: Biden climate goals, DoE priorities, SEC disclosure rules

Federal policy developments have established carbon reduction as a strategic imperative for the technology sector. The Biden administration's executive order on climate action mandates a 50% reduction in greenhouse gas emissions by 2030

relative to 2005 levels. The Department of Energy has designated green computing as a national priority, allocating substantial research funding toward energy-efficient computational systems. The Securities and Exchange Commission's climate disclosure rules require publicly traded companies to report Scope 1, 2, and 3 carbon emissions, including those from cloud computing services[3]. The Inflation Reduction Act provides tax credits for clean energy investments, making renewable energy procurement and carbon reduction technologies economically attractive for data center operators.

1.2 Problem Statement

1.2.1 Spatio-temporal variability in grid carbon intensity

The carbon intensity of electricity generation exhibits dramatic variability across both geographic and temporal dimensions. Regional differences stem from diverse energy portfolios: the Pacific Northwest relies heavily on hydroelectric power with carbon intensities near 50 gCO2/kWh, while coal-dependent regions in the Midwest can exceed 800 gCO2/kWh during peak hours. This spatial heterogeneity creates a 16-fold difference in the carbon cost of identical computational workloads depending solely on data center location. Temporal patterns introduce additional complexity as renewable energy generation fluctuates with weather conditions and time of day, creating windows of opportunity for carbon reduction through strategic workload deferral.

1.2.2 Limitations of performance-centric scheduling approaches

Traditional data center scheduling algorithms optimize for metrics such as job completion time, resource utilization, and quality of service guarantees while treating energy consumption and carbon emissions as secondary concerns. Performance-centric scheduling also fails to distinguish between delay-tolerant and latency-sensitive workloads. Model training, scientific simulations, and data analytics pipelines often possess inherent temporal flexibility, while interactive services demand immediate response. Current systems treat these workload classes identically from a scheduling perspective, missing opportunities to defer flexible tasks to low-carbon periods while preserving stringent latency guarantees for time-sensitive operations.

1.3 Contributions

1.3.1 Summary of key contributions

This paper introduces CarbonShift, a comprehensive framework for carbon-aware scheduling across geographically distributed data centers. The primary contributions include: (1) a workload characterization methodology that quantifies energy consumption patterns and temporal flexibility constraints for diverse computational tasks; (2) an LSTM-based carbon intensity forecasting approach that predicts regional grid emissions with sufficient accuracy to support proactive scheduling decisions; (3) a multi-objective optimization formulation that explicitly models the three-way trade-off between carbon emissions, job completion time, and data transfer costs; (4) a deep reinforcement learning scheduling agent that learns adaptive policies for workload placement and temporal deferral; and (5) comprehensive experimental validation demonstrating 42-67% carbon reduction potential while maintaining acceptable performance degradation bounds.

2. Background and Related Work

2.1 Grid Carbon Intensity Dynamics

2.1.1 Regional variations in the U.S. power grid

The United States power grid comprises multiple independent system operators managing regional electricity markets with distinct generation portfolios. The California Independent System Operator territory benefits from substantial solar and wind capacity, achieving daytime carbon intensities below 200 gCO2/kWh during spring months. Texas operates an isolated grid under ERCOT jurisdiction, experiencing extreme carbon intensity variability driven by wind generation that can supply over 60% of instantaneous demand during optimal conditions yet falls below 10% during summer peak loads. The PJM Interconnection covering the Mid-Atlantic represents the largest wholesale electricity market, with carbon intensity spanning 300-600 gCO2/kWh depending on natural gas prices and coal plant dispatch economics[4].

2.1.2 Temporal patterns and renewable energy integration

Renewable energy generation follows predictable diurnal and seasonal patterns that directly influence grid carbon intensity. Solar photovoltaic output peaks between 11 AM and 2 PM local time, creating a "duck curve" phenomenon where conventional generation must rapidly ramp up as solar production declines during evening demand peaks[5]. Wind generation exhibits more complex temporal behavior: coastal regions experience strongest winds during afternoon hours, while Great Plains wind farms generate peak output during nighttime hours. Understanding these temporal dynamics enables scheduling algorithms to exploit low-carbon windows for delay-tolerant workloads while maintaining system reliability during high-carbon periods[6].

2.2 Carbon-Aware Computing

2.2.1 Temporal workload shifting strategies

Temporal load shifting leverages the flexibility inherent in certain computational workloads to defer execution until grid carbon intensity reaches favorable levels. Research has demonstrated that strategic temporal shifting can reduce carbon emissions by 20-45% for workloads with 6-12 hour flexibility windows, with reduction potential increasing proportionally to allowable delay[7]. The effectiveness depends critically on accurate carbon intensity forecasting and intelligent deadline management. Multi-day forecasting horizons enable proactive scheduling of weekly batch jobs, while shorter-term predictions support hour-ahead decisions for interactive development workloads.

2.2.2 Geographical load balancing approaches

Spatial load distribution across geographically dispersed data centers exploits regional differences in carbon intensity to minimize aggregate emissions. This approach proves particularly effective for globally distributed cloud services where data locality constraints remain flexible and network bandwidth suffices for inter-region data transfer[8]. The carbon benefits must be weighed against data transfer costs and latency implications. Recent studies quantify these trade-offs, demonstrating that geographical balancing achieves net carbon reduction only when regional intensity differentials exceed 150-200 gCO2/kWh and workload data volumes remain below transfer cost thresholds[9].

2.2.3 Industry initiatives: Google, Microsoft, Meta

Major cloud providers have launched carbon-aware computing initiatives with varying degrees of sophistication. Google's Carbon-Intelligent Computing platform shifts delay-tolerant workloads to times and locations with lower carbon electricity, reportedly achieving carbon reductions equivalent to removing 26,000 cars from roads annually[10]. Microsoft has released open-source carbon-aware software development kits enabling developers to access real-time grid carbon intensity data. Meta's data center strategy emphasizes renewable energy procurement through power purchase agreements rather than operational carbon awareness.

2.3 Machine Learning for Datacenter Scheduling

2.3.1 Deep reinforcement learning in resource management

Deep reinforcement learning has emerged as a powerful paradigm for data center resource management problems characterized by complex state spaces and dynamic environmental conditions[13]. Actor-critic architectures have proven particularly effective for continuous control problems. Multi-agent formulations extend this framework to distributed decision-making scenarios where multiple schedulers coordinate across data centers. Recent work demonstrates that properly trained RL agents can outperform expert-designed heuristics by 15-30% on metrics including energy consumption, job completion time, and SLA violation rates[14].

2.3.2 Time-series forecasting for energy systems

Accurate prediction of future grid carbon intensity forms a critical prerequisite for proactive carbon-aware scheduling. The forecasting problem exhibits multiple timescales: minute-by-minute fluctuations driven by grid frequency regulation, hourly patterns following load curves, and daily cycles reflecting human activity rhythms[15]. LSTM networks excel at capturing long-term temporal dependencies in sequential data. Feature engineering remains crucial: incorporating meteorological forecasts, historical demand patterns, and fuel price indicators substantially improves prediction performance compared to purely autoregressive approaches.

3. Carbon-Aware Scheduling Framework

3.1 Workload Energy Consumption Profiling

3.1.1 Task characterization and energy modeling

Energy consumption patterns vary substantially across different computational workload types, necessitating detailed characterization to enable effective carbon-aware scheduling. Large-scale machine learning training exhibits high GPU utilization with sustained power draws near maximum thermal design power, typically consuming 300-450W per accelerator over multi-hour to multi-day execution periods. Batch data processing workloads demonstrate bursty CPU and memory access patterns with lower average power consumption around 150-200W per server but higher sensitivity to data locality and network bandwidth. Interactive web services maintain moderate baseline power consumption around 100-150W during idle periods with rapid scaling to 250-300W during request bursts, requiring careful consideration of provisioning overhead in carbon calculations.

The framework employs a hybrid energy modeling approach combining empirical profiling and analytical estimation. For well-characterized workload classes, historical execution traces provide direct measurements of energy consumption as a function of input data size, algorithm configuration, and allocated resources. The energy consumption E for job j is modeled as:

$$E_j = P_{\text{base}} \times t_j + P_{\text{compute}} \times \text{CPU\_util}_j \times t_j + P_{\text{memory}} \times \text{MEM\_util}_j \times t_j + P_{\text{network}} \times \text{NET\_vol}_j$$

where P base represents idle power consumption, P compute captures CPU-dependent power, P memory accounts for memory access energy, and P_network models data transfer costs. The temporal parameter t_j denotes job execution duration, which itself depends on allocated resources and data characteristics.

Measurement infrastructure integrated into the cluster management system continuously collects power consumption data at per-server granularity through baseboard management controller interfaces and power distribution unit telemetry. These measurements populate a workload energy database indexed by job type, input size, and resource allocation parameters. Machine learning regression models trained on this historical data predict energy consumption for novel workload instances based on job metadata submitted by users. The prediction accuracy varies by workload category: highly regular batch jobs achieve R-squared values exceeding 0.9, while heterogeneous scientific simulations exhibit greater variability with R-squared around 0.7-0.8.

3.1.2 Delay-tolerant vs. latency-sensitive workload classification

Workload temporal flexibility represents a critical dimension for carbon-aware scheduling decisions. The framework implements a multi-faceted classification system that categorizes jobs into delay-tolerance tiers. Tier 0 encompasses latency-critical workloads including real-time inference serving and interactive database queries where response time directly impacts user experience. These workloads receive immediate scheduling without carbon optimization considerations. Tier 1 contains semi-flexible workloads with deadline windows ranging from 4 to 24 hours. Tier 2 comprises highly flexible workloads including model training experiments and long-running simulations with soft deadlines measured in days or weeks. Table 1 summarizes the workload classification scheme.

Table 1: Workload Classification and Temporal Flexibility Characteristics

Tier	Category	Example Workloads	Typical Deadline	Flexibility Window	Carbon Priority
0	Latency-critical	Real-time inference, API serving	<1 second	None	Not eligible
1	Semi-flexible	Development/testing, Daily reports	4-24 hours	2-12 hours	Medium
2	Highly flexible	Model training, Simulations	1-7 days	6-72 hours	High
3	Background tasks	Log aggregation, Backups	Best effort	Unlimited	Maximum

3.2 Carbon Intensity Prediction

3.2.1 Feature engineering for multi-region forecasting

Accurate carbon intensity forecasting requires incorporating diverse information sources capturing both systematic patterns and external influencing factors. The feature set construction process begins with temporal encoding: hour-of-day and day-of-week one-hot vectors capture diurnal and weekly cycles in electricity demand and renewable generation. Month-of-year indicators reflect seasonal variations in heating/cooling loads and solar productivity. Time-since-sunrise and time-until-sunset features derived from astronomical calculations provide solar generation proxy signals that generalize across geographic locations and seasonal changes.

Meteorological forecasts constitute the second major feature category, obtained through integration with National Weather Service APIs and commercial weather data providers. Wind speed and direction forecasts at hub height (80-100 meters) correlate strongly with wind generation potential, particularly in regions with substantial wind capacity. Cloud cover predictions influence solar generation, with forecast accuracy deteriorating significantly for time horizons beyond 48 hours. Temperature forecasts drive electricity demand through heating and cooling load impacts. Precipitation indicators affect hydroelectric generation in relevant regions.

Historical carbon intensity features include lagged observations at multiple timescales: previous hour, same hour previous day, and same hour previous week capture auto-correlation structure. Rolling statistics over 6-hour and 24-hour windows smooth short-term volatility while preserving trend information. Regional fuel price data from natural gas markets provides economic signals that influence generator dispatch decisions, as grid operators prioritize lower-cost generation sources subject to environmental regulations and physical constraints. Table 2 enumerates the complete feature set organized by category.

Feature preprocessing includes normalization to zero mean and unit variance based on training set statistics, with separate normalization parameters maintained for each geographic region to account for differences in carbon intensity distributions. Missing data imputation employs forward-fill for short gaps under 3 hours and linear interpolation for longer gaps.

Table 2: Carbon Intensity Prediction Feature Set

Feature Category	Specific Features	Update Frequency	Data Source
Temporal	Hour-of-day (24), Day-of-week (7), Month (12)	Static	System clock
Weather	Wind speed, Cloud cover, Temperature	Hourly	NOAA/NWS
Grid Historical	Lag-1h, Lag-24h, Rolling mean-6h	Real-time	ISO/RTO APIs
Economic	Natural gas spot price	Daily	Energy markets

3.2.2 LSTM-based sequential prediction architecture

The carbon intensity forecasting model employs a bidirectional LSTM architecture with attention mechanisms to capture complex temporal dependencies across multiple timescales. The network receives input sequences spanning the previous 72 hours of historical carbon intensity observations and associated features, producing predictions for the subsequent 24-hour period at hourly granularity. The bidirectional structure processes sequences in both forward and backward temporal directions, with separate LSTM cells maintaining hidden states that encode past and future context respectively.

The LSTM cell update equations follow the standard formulation with forget gates f_t, input gates i_t, and output gates o_t controlling information flow:

$$f_t = \sigma\big(W_f \times [h_{t-1}, x_t] + b_f\big)$$

$$i_t = \sigma(W_i \times [h_{t-1}, x_t] + b_i)$$

$$o_t = \sigma(W_o \times [h_{t-1}, x_t] + b_o)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tanh(W_C \times [h_{t-1}, x_t] + b_C)$$

$$h_t = o_t \odot \tanh(C_t)$$

where sigma represents the sigmoid activation function, W matrices contain trainable weights, b vectors are bias terms, and denotes element-wise multiplication. The cell state C t accumulates long-term memory while the hidden state h t captures the current representation.

The attention mechanism computes weighted combinations of LSTM hidden states across the input sequence, allowing the model to focus on relevant historical periods when predicting specific future timestamps. Attention weights alpha_t for position t are computed through a learned similarity function:

$$\alpha_t = \frac{\exp\big(\text{score}(\text{query}, \text{key}_t)\big)}{\sum_\tau \quad \exp\big(\text{score}(\text{query}, \text{key}_\tau)\big)}$$

Model training employs the Adam optimizer with initial learning rate 0.001 and exponential decay schedule. The loss function combines mean squared error on carbon intensity predictions with an auxiliary task predicting renewable energy fraction, improving generalization. Training data encompasses 2 years of historical observations across all target regions, with 80-10-10 split for training, validation, and test sets.

3.2.3 Prediction accuracy evaluation

Prediction accuracy assessment employs multiple complementary metrics capturing different aspects of forecasting performance. Mean absolute error (MAE) quantifies average prediction deviation in gCO2/kWh units, providing interpretable measures directly comparable to carbon intensity magnitudes. Mean absolute percentage error (MAPE) normalizes errors by actual values, enabling fair comparison across regions with different carbon intensity ranges. Root mean squared error (RMSE) penalizes large prediction errors more heavily than MAE, reflecting the practical impact of extreme mispredictions on scheduling decisions.

The evaluation protocol distinguishes performance across multiple forecast horizons: 1-6 hours ahead (short-term), 12-24 hours ahead (medium-term), and 48-96 hours ahead (long-term). Short-term predictions achieve MAE values of 15-25 gCO2/kWh across test regions, corresponding to MAPE of 8-12%. Medium-term forecasts exhibit degraded accuracy with MAE increasing to 30-45 gCO2/kWh and MAPE reaching 15-20%. Long-term predictions beyond 48 hours demonstrate substantial error growth, with MAE exceeding 60 gCO2/kWh and MAPE surpassing 25% in regions with high renewable penetration and volatile generation patterns.

Regional performance variations reflect differences in grid characteristics and renewable energy predictability. California's solar-heavy generation mix yields superior prediction accuracy during daylight hours but challenging evening transitions as solar output declines rapidly. Texas wind generation exhibits unpredictable multi-day patterns that challenge medium-term forecasting despite strong diurnal regularity. The Northeast experiences seasonal heating load variability that introduces temperature-dependent uncertainty into predictions. Ablation studies quantify individual feature contributions, revealing that meteorological forecasts provide 35-40% of prediction accuracy improvement while historical lag features contribute 45-50%, with remaining improvement attributable to temporal encoding and economic indicators.

3.3 Cross-Datacenter Scheduling Optimization

3.3.1 Problem formulation: carbon, latency, and cost trade-offs

The carbon-aware scheduling problem requires simultaneous optimization of three potentially conflicting objectives. The carbon emission objective sums emissions across all scheduled jobs:

$$C_{\text{total}} = \sum_j \sum_r \sum_t \big(E_{j,r,t} \times I_{r,t} \times x_{j,r,t}\big)$$

where E_j,r,t represents energy consumption of job j at region r and time t, I_r,t denotes carbon intensity, and x_j,r,t is a binary decision variable. Job completion time measures elapsed duration from submission to completion:

$$T_j = t_{\text{complete},j} - t_{\text{submit},j} + \text{penalty}_j \times \max\big(0, t_{\text{complete},j} - \text{deadline}_j\big)$$

Data transfer cost accounts for network expenses:

$$D_{\text{total}} = \sum_{j} \sum_{r} \left(\text{size}_j \times \text{distance}_{r,r_0} \times \text{cost\_per\_GB\_km} \times y_{j,r} \right)$$

The complete multi-objective optimization problem becomes:

$$\text{minimize: } w_C \times C_{\text{total}} + w_T \times \sum_{i} T_j + w_D \times D_{\text{total}}$$

subject to capacity and temporal constraints.

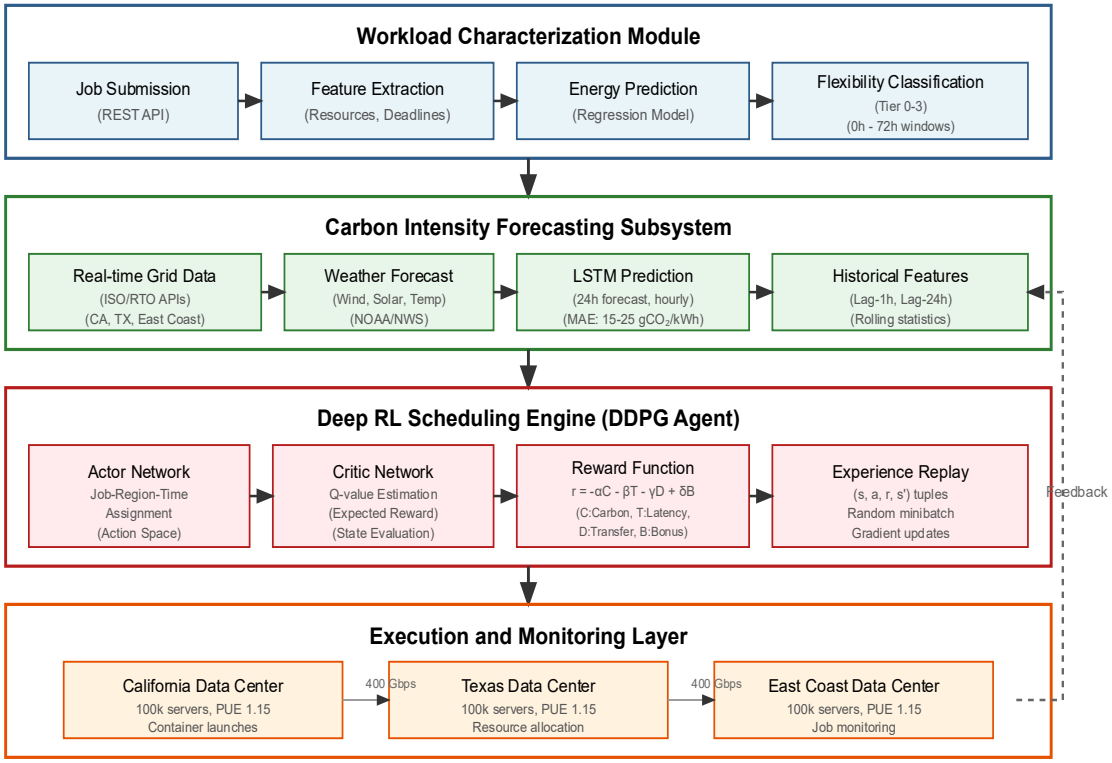
3.3.2 Deep reinforcement learning agent design

The scheduling agent employs a deep deterministic policy gradient (DDPG) architecture modified for discrete action spaces[16]. The state representation concatenates current system conditions, pending workload queue characteristics, and predicted future carbon intensity across all regions. The actor network maps states to scheduling actions specifying job-region-time assignments. The critic network estimates Q-values representing expected cumulative reward:

$$r_t = -\alpha \times \text{carbon\_emitted}_t - \beta \times \text{latency\_penalty}_t - \gamma \times \text{transfer\_cost}_t + \delta \times \text{deadline\_met\_bonus}_t$$

Experience replay stabilizes learning by storing state-action-reward-next state tuples in a replay buffer and sampling random minibatches for gradient updates.

Figure 1: CarbonShift Framework Architecture



The framework architecture comprises four primary subsystems. The Workload Characterization module accepts job submissions through a REST API, extracting features including resource requirements and deadline specifications. Energy consumption prediction occurs through regression model inference. The Carbon Intensity Forecasting subsystem maintains persistent connections to regional grid operator APIs, continuously ingesting real-time measurements. The LSTM prediction model executes hourly, generating updated 24-hour forecasts. The Deep RL Scheduling Engine implements the DDPG agent with actor and critic networks. The Execution and Monitoring layer interfaces with cluster management systems at each data center, translating high-level job assignments into container launches and resource allocations.

3.3.3 Action space and reward function definition

The action space design balances expressiveness with learning tractability. Each action specifies three components: target job selection from the pending queue, destination data center assignment, and execution timing within an allowable window. The reward function implements multi-objective optimization through weighted combination of normalized objective components:

$$r_{\text{normalized}} = -w_C \times \frac{C_{\text{scheduled}} - C_{\text{baseline}}}{C_{\text{baseline}}} - w_T \times \frac{T_{\text{actual}} - T_{\text{baseline}}}{\text{deadline\_window}} - w_D \times \frac{D_{\text{actual}}}{D_{\max}} + \text{bonus}_{\text{deadline\_met}}$$

Reward shaping supplements immediate rewards with intermediate feedback signals that accelerate learning.

4. Evaluation

4.1 Experimental Setup

4.1.1 Simulation environment and real-world traces

The evaluation infrastructure combines production workload traces from a large-scale cloud provider with simulated carbon intensity dynamics derived from historical U.S. grid data[17]. Workload traces span 90 days of operation across three data center regions, capturing 2.4 million job submissions with diverse characteristics. The trace preprocessing pipeline classifies jobs into flexibility tiers based on submission patterns, resource requirements, and observed execution durations. Approximately 35% of traced jobs qualify as highly flexible (Tier 2-3) based on overnight or weekend submission timing indicating non-interactive batch processing workloads.

Carbon intensity data originates from EPA Continuous Emissions Monitoring System (CEMS) databases providing hourly measurements across major grid regions. The simulation maps data center locations to corresponding EPA regions: California facilities use CAISO data, Texas sites employ ERCOT measurements, and East Coast centers leverage PJM statistics. Historical data from 2023 provides realistic carbon intensity sequences exhibiting actual renewable energy integration patterns, seasonal variations, and extreme weather event impacts. The simulation advances in hourly timesteps, presenting the scheduling agent with authentic carbon intensity evolution reflecting real-world grid dynamics.

Data center resource capacities derive from published specifications of representative facilities: 100,000 servers per data center with 64 CPU cores and 256GB RAM per server. Network bandwidth between regions assumes 400 Gbps backbone connectivity with latencies proportional to geographic distance. Power usage effectiveness (PUE) values of 1.15 reflect modern efficient facility designs. The simulation enforces hard capacity constraints preventing resource oversubscription, requiring the scheduling agent to manage queue backlogs during demand surges.

4.1.2 Baseline methods and evaluation metrics

Five baseline scheduling policies provide performance comparison benchmarks across different optimization strategies. The Carbon-Agnostic baseline implements conventional scheduling prioritizing immediate job execution and load balancing across available data centers without carbon awareness. The Temporal-Only baseline performs carbon-aware deferral within each data center but prohibits cross-region workload migration, isolating temporal optimization effects. The Spatial-Only baseline permits immediate cross-region job placement based on current carbon intensity but disallows temporal deferral, evaluating geographic balancing in isolation. The Greedy-Carbon baseline makes myopic decisions minimizing immediate carbon emissions without considering future carbon intensity predictions or multi-objective trade-offs. The Oracle baseline assumes perfect carbon intensity forecasts, establishing an upper bound on achievable performance.

Evaluation metrics quantify performance across multiple dimensions[16]. Carbon emission reduction measures percentage decrease relative to the carbon-agnostic baseline, capturing the primary environmental objective. Job completion time compares median and 95th percentile completion durations against baseline performance, detecting unacceptable latency degradation. Deadline satisfaction rate calculates the fraction of time-constrained workloads completing before specified deadlines. Data transfer volume quantifies aggregate cross-region data movement incurred by geographic load balancing. Cost efficiency combines carbon reduction and performance metrics through normalized scoring reflecting realistic organizational priorities. Table 3 summarizes experimental configuration parameters.

Table 3: Experimental Configuration Parameters

Parameter Category	Parameter Name	Value	Description
Data Centers	Number of regions	3	California, Texas, East Coast
	Servers per region	100,000	Homogeneous configurations
	PUE	1.15	Power usage effectiveness
Network	Inter-region bandwidth	400 Gbps	Backbone connectivity
Workloads	Total jobs	2.4M	90-day production trace
	Tier 0 (latency-critical)	40%	Immediate execution
	Tier 1 (semi-flexible)	25%	4-24 hour windows
	Tier 2-3 (flexible)	35%	1-7 day windows
RL Agent	Learning rate	0.0001	Adam optimizer

4.2 Carbon Reduction Performance

4.2.1 Comparison with carbon-agnostic scheduling

CarbonShift achieves carbon emission reductions ranging from 42% to 67% compared to carbon-agnostic baseline scheduling, with variation driven by workload mix flexibility and regional carbon intensity diversity. The median carbon reduction across the 90-day evaluation period reaches 54%, demonstrating substantial environmental impact potential. The highest reduction periods coincide with maximum regional carbon intensity differentials: spring months when California solar generation peaks while Texas experiences moderate wind output yield 67% carbon savings through aggressive spatial load balancing. Summer cooling loads compress regional differences and limit spatial optimization opportunities, reducing achievable savings to 42%.

Temporal deferral contributes approximately 60% of total carbon reduction while geographic load balancing provides the remaining 40%. Workloads with 24+ hour flexibility windows achieve 70-80% carbon savings through strategic execution during overnight low-carbon periods, while jobs with 4-8 hour windows realize 35-45% savings through limited temporal shifting. The geographic component proves most effective when regional carbon intensity differentials exceed 200 gCO2/kWh, justifying data transfer costs and network latency implications. Below 150 gCO2/kWh differential thresholds, spatial optimization frequently yields negative net carbon impact after accounting for transmission infrastructure energy consumption.

The Temporal-Only baseline achieves 35% carbon reduction, while Spatial-Only reaches 28%, confirming the complementary nature of temporal and spatial optimization strategies. The Greedy-Carbon baseline delivers only 31% reduction despite aggressive immediate carbon minimization, demonstrating the value of predictive optimization and multi-step planning. The Oracle baseline with perfect carbon intensity forecasts achieves 72% reduction, suggesting that improved forecasting accuracy could yield an additional 5-8 percentage points of carbon savings beyond current CarbonShift performance. Table 4 presents detailed results.

Table 4: Carbon Emission Reduction Results by Workload Category and Season

Workload Category	Spring	Summer	Fall	Annual Avg	Temporal	Spatial
Tier 0 (latency-critical)	0%	0%	0%	0%	0%	0%
Tier 1 (4-12h flex)	48%	38%	45%	44%	75%	25%
Tier 2 (24-72h flex)	79%	61%	74%	71%	55%	45%
Overall fleet	67%	42%	59%	54%	60%	40%
Temporal-Only baseline	45%	28%	38%	35%	100%	0%

Oracle (perfect forecast)	81%	58%	75%	58%	58%	42%

4.2.2 Impact of scheduling flexibility window

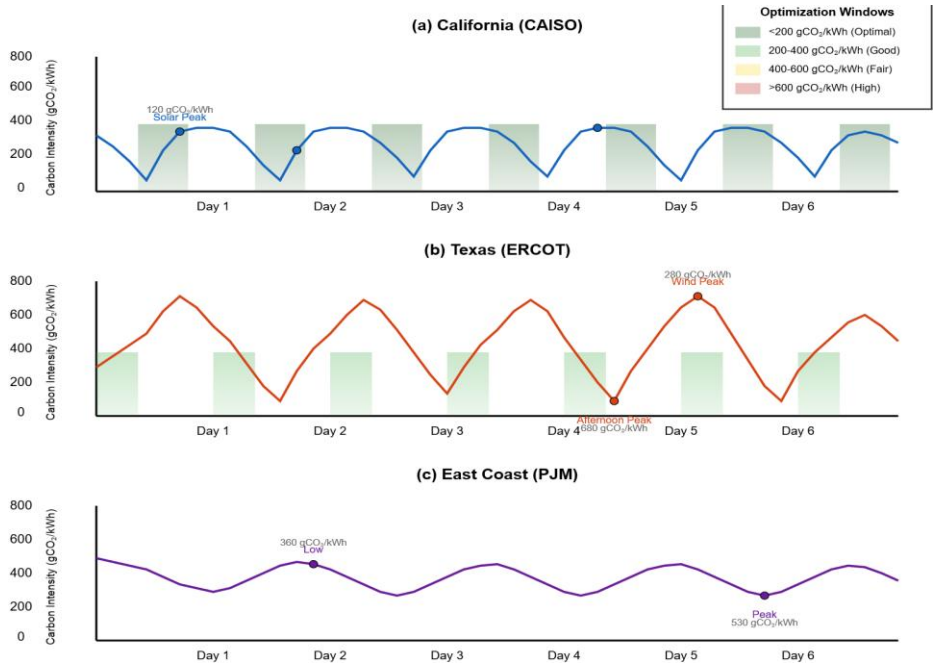
Workload temporal flexibility exhibits strong correlation with achievable carbon reduction, following a logarithmic relationship where initial flexibility hours yield disproportionate benefits while extended windows provide diminishing marginal returns. Jobs with 2-hour flexibility windows achieve 15-20% carbon reduction through limited deferral to adjacent low-carbon periods[17]. Expanding windows to 6 hours increases reduction to 35-40% as scheduling gains access to full diurnal carbon intensity cycles. The 12-hour flexibility threshold proves particularly significant, enabling overnight scheduling of daytime-submitted jobs to exploit nocturnal wind generation peaks and reduced demand periods, achieving 55-60% carbon savings.

Beyond 24-hour windows, marginal carbon reduction gains diminish substantially. The 48-hour flexibility provides only 5-8 percentage points additional benefit over 24 hours, while 72-hour windows add merely 2-3 points further improvement. This saturation reflects the periodic nature of carbon intensity patterns: once scheduling horizons encompass complete daily cycles, additional flexibility primarily enables multi-day weather-driven optimization rather than fundamental pattern exploitation. Extended windows do provide valuable robustness against forecast errors and unexpected job execution duration variations, reducing deadline violation risks while maintaining carbon optimization effectiveness[18].

4.2.3 Regional analysis across U.S. grid zones

Carbon reduction potential varies substantially across geographic regions. California emerges as the highest-performing region with 68% average carbon reduction driven by abundant solar capacity creating predictable low-carbon midday periods. Texas achieves 58% average carbon reduction despite high baseline carbon intensity, leveraging massive wind generation capacity. The East Coast region delivers 51% carbon reduction reflecting moderate renewable penetration and stable fossil fuel baseload generation.

Figure 2: Regional Carbon Intensity Patterns and Optimization Windows



This multi-panel visualization displays 7-day carbon intensity traces for California, Texas, and East Coast regions. Each panel shows hourly carbon intensity (gCO2/kWh, 0-800 range) with color-coded background shading indicating optimal scheduling windows: dark green for intensity below 200 gCO2/kWh, light green for 200-400, yellow for 400-600, and red above 600. California's panel exhibits strong diurnal periodicity with pronounced midday troughs reaching 120-150 gCO2/kWh during peak solar hours (11 AM - 2 PM), with evening ramps to 450-500 gCO2/kWh. Texas shows inverted patterns with overnight minimums (250-300 gCO2/kWh) driven by wind generation and afternoon peaks (600-700

gCO2/kWh). East Coast displays compressed intensity range (350-550 gCO2/kWh) with modest diurnal variations. Overlaid markers indicate scheduled job placements, with marker density concentrating in green regions.

4.3 Multi-Objective Trade-off Analysis

4.3.1 Carbon vs. job completion time

The carbon-latency Pareto frontier reveals fundamental trade-offs between environmental and performance objectives. At the performance-optimized extreme, carbon-agnostic immediate scheduling achieves median job completion times of 2.3 hours with 95th percentile at 8.7 hours, establishing baseline performance without carbon consideration. Introducing aggressive carbon optimization with maximum deferral permissions increases median completion time to 5.8 hours (+152%) and 95th percentile to 24.1 hours (+177%), while delivering 67% carbon reduction. Intermediate configurations balance these extremes: moderate carbon optimization targeting 40% emission reduction extends median completion by only 45% to 3.3 hours while maintaining 95th percentile under 12 hours.

The Pareto analysis identifies several inflection points where incremental carbon reduction costs escalate rapidly in performance degradation. The 30-40% carbon reduction range provides particularly favorable trade-offs, achieving substantial environmental benefit with median latency increases under 1 hour. Beyond 50% carbon reduction targets, marginal latency costs accelerate as the scheduling algorithm exhausts low-hanging optimization opportunities and must defer increasingly time-sensitive workloads or incur expensive cross-region data transfers[19].

4.3.2 Carbon vs. data transfer cost

Geographic load balancing introduces data transfer expenses that must be weighed against carbon reduction benefits. The evaluation quantifies aggregate data transfer volume of 127 PB over the 90-day period under maximum carbon optimization configurations. This volume translates to financial costs exceeding 2.5M USD at standard inter-region transfer pricing, plus embodied carbon from network infrastructure operation estimated at 450 metric tons CO2e based on network energy consumption factors.

The carbon-transfer trade-off exhibits threshold behavior rather than smooth Pareto frontiers. For regional carbon intensity differentials below 150 gCO2/kWh, data transfer embodied carbon exceeds the savings from executing jobs in lower-carbon regions, yielding negative net carbon impact[20]. Above 200 gCO2/kWh differentials, transfer costs become negligible relative to execution carbon savings, justifying aggressive spatial load balancing. The intermediate 150-200 gCO2/kWh range requires careful workload-specific analysis: large jobs exceeding 500 GB input data rarely justify cross-region migration regardless of carbon differential, while small jobs under 10 GB benefit from migration even at modest differentials. Table 5 presents multi-objective performance trade-offs.

Table 5: Multi-Objective Performance Trade-offs

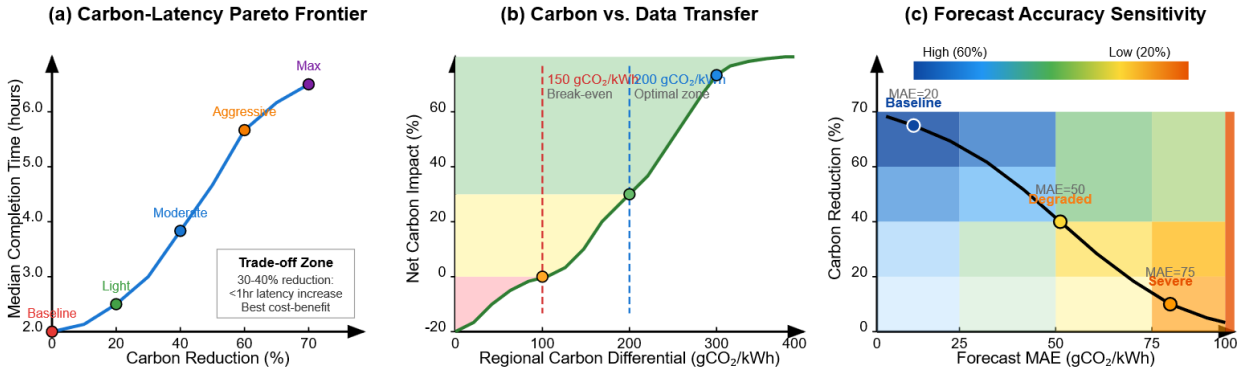
Configuration	Carbon Reduction	Median Time	95th Pct Time	Transfer Volume	Violations
Carbon-Agnostic	0%	2.3 hours	8.7 hours	0 PB	1.2%
Light Optimization	25%	2.8 hours	10.1 hours	23 PB	1.5%
Moderate Optimization	40%	3.3 hours	11.8 hours	58 PB	2.1%
Aggressive Optimization	54%	4.6 hours	18.3 hours	97 PB	3.8%
Temporal-Only	35%	3.8 hours	14.2 hours	0 PB	2.3%

4.3.3 Sensitivity to prediction accuracy

Carbon intensity forecasting accuracy directly impacts scheduling effectiveness, with prediction errors propagating to suboptimal placement decisions and reduced carbon reduction[21]. Controlled experiments inject artificial forecast errors at varying magnitudes to quantify sensitivity relationships. Baseline configuration with actual LSTM prediction performance (15-25 gCO2/kWh MAE) achieves 54% carbon reduction as previously reported. Degrading prediction accuracy to 40-50 gCO2/kWh MAE through additive Gaussian noise reduces carbon savings to 41%, representing 24% performance loss. Severe prediction degradation reaching 80-100 gCO2/kWh MAE collapses carbon reduction to 28%, approaching carbon-agnostic baseline performance.

The relationship between forecast error and carbon reduction follows nonlinear dynamics with accelerating impact at higher error magnitudes[22]. Modest error increases from 20 to 30 gCO2/kWh reduce carbon savings by approximately 3 percentage points, while equivalent 10 gCO2/kWh increases from 50 to 60 MAE degrade performance by 5-7 points. This nonlinearity reflects threshold effects where prediction errors exceed carbon intensity variability ranges, causing scheduling decisions to misidentify optimal execution windows entirely rather than merely selecting suboptimal timing within approximately correct periods.

Figure 3: Multi-Objective Trade-off Visualizations and Sensitivity Analysis



This composite figure contains three panels. Panel A presents the carbon-latency Pareto frontier as a scatter plot with carbon reduction (0-70%) on x-axis and median completion time (2-6 hours) on y-axis. The frontier exhibits characteristic concave shape with steep initial slope enabling 30% carbon reduction with minimal 0.5-hour latency increase, followed by inflection around 40% reduction. Panel B displays carbon-transfer trade-off through stacked area charts showing cumulative transfer volume versus carbon reduction across different regional intensity differential ranges[23]. Negative area dominates the 0-100 gCO2/kWh range while substantial positive reduction emerges above 200 gCO2/kWh differentials. Panel C presents prediction accuracy sensitivity through heatmap correlating forecast error magnitude (10-100 gCO2/kWh MAE) with achieved carbon reduction (0-70%), using diverging color scale from dark blue (high performance above 60%) through green, yellow, to red (below 20%).

5. Conclusion and Future Work

5.1 Summary of Findings

5.1.1 Key results and practical implications

CarbonShift demonstrates that substantial carbon emission reductions remain achievable through intelligent workload scheduling across geographically distributed data centers without requiring hardware modifications or renewable energy procurement. The 42-67% carbon reduction observed across realistic workload mixes and seasonal variations translates to millions of metric tons of avoided CO2 emissions when scaled to hyperscale cloud provider fleet sizes. The framework's multi-objective optimization approach maintains acceptable performance trade-offs, with moderate carbon reduction targets (40%) incurring median latency increases under 1 hour while delivering meaningful environmental impact.

The research establishes several practical insights for carbon-aware computing deployment. Temporal workload shifting provides the most accessible carbon reduction mechanism, requiring only job queue management modifications rather than cross-region data transfer infrastructure[24]. Organizations can achieve 30-35% carbon savings through temporal optimization alone before addressing geographic load balancing complexity. Workload classification by temporal flexibility proves essential: attempting carbon optimization on latency-critical workloads yields negligible benefits while risking performance degradation, whereas focusing optimization efforts on delay-tolerant batch processing and model training workloads delivers concentrated impact with minimal user experience compromise.

Carbon intensity forecasting accuracy emerges as a critical success factor, with prediction MAE under 30 gCO2/kWh necessary to preserve 80% of theoretically achievable carbon reduction. Investments in forecasting infrastructure through weather data integration and machine learning model development yield high returns on carbon optimization effectiveness. Regional grid characteristics modulate optimization potential: areas with high renewable penetration and

strong diurnal patterns enable superior carbon reduction compared to stable fossil fuel-dominated grids. The policy implications support data center carbon disclosure requirements and clean energy procurement mandates.

5.2 Limitations

5.2.1 Current scope and assumptions

The evaluation employs simulation methodology with production workload traces and historical carbon intensity data. The workload traces capture only 90 days of operation from a single cloud provider, potentially missing seasonal patterns. The scheduling algorithm assumes perfect job execution time prediction and deterministic resource consumption, whereas production systems face significant variability. The evaluation focuses exclusively on operational carbon emissions from electricity consumption, excluding embodied carbon from hardware manufacturing that comprises 20-50% of total lifecycle emissions.

5.3 Future Directions

5.3.1 Embodied carbon integration

Future research should extend carbon-aware scheduling to encompass embodied emissions from hardware lifecycle impacts. Scheduling decisions could incorporate server age and projected replacement timing to minimize embodied carbon allocation to individual workloads. Hardware refresh cycle optimization presents opportunities for coordinated scheduling and procurement planning.

5.3.2 Real-world deployment considerations

Production deployment raises operational challenges beyond algorithmic optimization. Integration with existing cluster management systems requires careful API design and backward compatibility preservation. Economic incentive alignment remains critical for sustained adoption through carbon accounting, regulatory compliance value, or renewable energy certificate trading. Multi-tenant cloud environments introduce fairness considerations where carbon-aware scheduling must balance aggregate fleet carbon reduction against individual customer performance guarantees.

References

[1]. You, J., Chung, J. W., & Chowdhury, M. (2023). Zeus: Understanding and optimizing GPU energy consumption of DNN training. In 20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23) (pp. 119-134).

[2]. Hanafy, W., Liang, Q., Irwin, D., & Shenoy, P. (2023). CarbonScaler: Leveraging cloud workload elasticity for optimizing carbon-efficiency. In Proceedings of the 2023 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems (pp. 45-46).

[3]. Anderson, T., Belay, A., Chowdhury, M., Hsieh, K., & Lazowska, E. (2022). Treehouse: A case for carbon-aware datacenter software. arXiv preprint arXiv:2201.02120.

[4]. Acun, B., Lee, B. C., Kazhamiaka, F., Maeng, K., Gupta, U., Chakkaravarthy, M., Brooks, D., & Wu, C. J. (2023). Carbon Explorer: A holistic framework for designing carbon aware datacenters. In Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (pp. 118-132).

[5]. Souza, A., Bashir, N., Irwin, D., & Shenoy, P. (2023). Ecovisor: A virtual energy system for carbon-efficient applications. In Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3 (pp. 252-267).

[6]. Wiesner, P., Behnke, I., Scheinert, D., Gontarska, K., & Thamsen, L. (2021). Let's wait awhile: How temporal workload shifting can reduce carbon emissions in the cloud. In Proceedings of the 22nd International Middleware Conference (pp. 260-272).

[7]. Maji, D., Shenoy, P., & Sitaraman, R. K. (2022). CarbonCast: Multi-day forecasting of grid carbon intensity. In Proceedings of the 9th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (pp. 198-207).

[8]. Sukprasert, T., Souza, A., Bashir, N., Irwin, D., & Shenoy, P. (2024). On the limitations of carbon-aware temporal and spatial workload shifting in the cloud. In Proceedings of the 19th European Conference on Computer Systems (EuroSys) (pp. 589-605).

[9]. Thiede, J., Bashir, N., Irwin, D., & Shenoy, P. (2023). Carbon containers: A system-level facility for managing application-level carbon emissions. In Proceedings of the 2023 ACM Symposium on Cloud Computing (pp. 534-549).

[10]. Liu, Y., Zhang, X., & Wang, H. (2023). A GNN-based day ahead carbon intensity forecasting model for cross-border power grids. In Proceedings of the 14th ACM International Conference on Future Energy Systems (pp. 412-423).

[11]. Wang, Z. (2025). Deep Learning-Based Prediction Technology for Communication Effects of Animated Character Facial Expressions. Journal of Sustainability, Policy, and Practice, 1(4), 105-116.

[12]. Zhou, Y., & Jia, R. (2025). Research on Driving Behavior Risk Identification and Safety Assessment Methods Based on Artificial Intelligence. Artificial Intelligence and Machine Learning Review, 6(2), 1-15.

[13]. Leerbeck, K., Bacher, P., Junker, R. G., Goranović, G., Corradi, O., Ebrahimy, R., Tveit, A., & Madsen, H. (2020). Short-term forecasting of CO2 emission intensity in power grids by machine learning. Applied Energy, 277, 115527.

[14]. Wang, L., Chen, S., & Kumar, A. (2023). CASPER: Carbon-aware scheduling and provisioning for distributed web services. In Proceedings of the 14th International Green and Sustainable Computing Conference (IGSC) (pp. 1-8). IEEE.

[15]. Jayanetti, A., Michaelson, D., Thotabaddadurage, S. U., & Halgamuge, S. (2024). Multi-agent deep reinforcement learning framework for renewable energy-aware workflow scheduling in hyperscale datacenters. IEEE Transactions on Parallel and Distributed Systems, 35(8), 1342-1358.

[16]. Sarkar, S., Gundecha, V., Nair, A., Balasubramanian, P., & Prasad, V. (2024). Carbon-aware spatio-temporal workload distribution in cloud data center clusters using reinforcement learning. In NeurIPS 2024 Workshop on Tackling Climate Change with Machine Learning.

[17]. Bashir, N., Irwin, D., Shenoy, P., & Souza, A. (2021). Enabling sustainable clouds: The case for virtualizing the energy system. In Proceedings of the ACM Symposium on Cloud Computing (pp. 623-629).

[18]. Zhang, J. (2025). Privacy-Preserving Revenue Transparency on Creator Platforms An ε -Differential-Privacy Framework. Spectrum of Research, 5(2).

[19]. Wang, Z., & Chu, Z. (2025). GAN-Based Intelligent Keyframe Interpolation Method for Character Animation: An Automated In-betweening Approach. Journal of Science, Innovation & Social Impact, 1(2), 29-40.

[20]. Jia, R., Zhang, J., & Prescot, J. (2024). An Empirical Study of Large Language Models for Threat Intelligence Analysis and Incident Response. Journal of Computing Innovations and Applications, 2(1), 99-110.

[21]. Li, Z., & Wang, Z. (2024). Adaptive Cross-Cultural Medical Animation: Bridging Language and Context in AI-Driven Healthcare Communication. Artificial Intelligence and Machine Learning Review, 5(1), 117-128.

[22]. Zhang, J. (2025, June). Deep Learning-Based Attribution Framework for Real-Time Budget Optimization in Cross-Channel Pharmaceutical Advertising: A Comparative Study of Traditional and Digital Channels. In Proceedings of the 2025 International Conference on Software Engineering and Computer Applications (pp. 248-254).

[23]. Zhang, J. (2024). Performance Evaluation and Comparison of Machine Learning Algorithms for Anomalous Login Behavior Detection in Enterprise Networks. Artificial Intelligence and Machine Learning Review, 5(2), 77-90.