LEARNING FROM MIXED-QUALITY DATA VIA EXPLORATION-GUIDED COGNITIVE MANIFOLDS

Anonymous authors
Paper under double-blind review

ABSTRACT

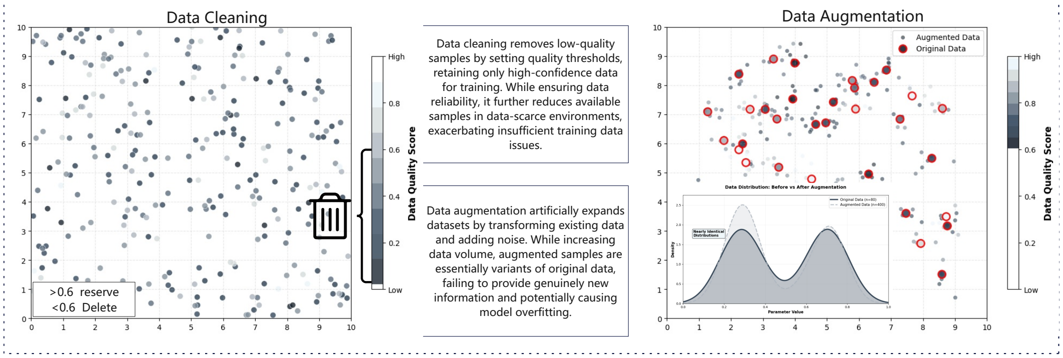
Data-driven learning paradigms typically require massive amounts of high-quality data to learn reliable decision models. In safety-critical domains, the cost of obtaining high-quality labeled data often exceeds practical budgets, while abundant low-quality data from routine operations remains unutilized. This creates a paradox: systems are data-rich but knowledge-poor, as traditional methods require extensive preprocessing that often removes valuable information. We introduce a learning framework that constructs cognitive manifolds to represent the system's evolving understanding, enabling principled utilization of mixed-quality data. Rather than filtering out imperfect data or treating all data equally, our approach dynamically evaluates each data point's potential contribution based on the learner's current cognitive state. We formalize this through exploration measures that capture the model's understanding uncertainty, enabling it to identify when low-quality data can fill knowledge gaps versus when it would introduce harmful noise. The framework transforms the traditional data quality problem into a cognitive matching problem: aligning data characteristics with the learner's current knowledge needs. Through this cognitive manifold representation, even highly noisy data becomes valuable during early learning stages when any information reduces uncertainty, while the same data is appropriately ignored once the model develops robust understanding. Experiments on multi-objective optimization tasks with mixed-quality data demonstrate that our approach extracts 3.6% more actionable knowledge from the same datasets compared to quality-filtering baselines, while maintaining the strict constraint satisfaction required in safety-critical applications.

1 INTRODUCTION

Mission-critical renewable energy power systems require intelligent dispatching strategies to achieve multi-objective optimisation whilst satisfying hard constraints (Rao et al., 2025). This challenge is particularly pronounced in satellite battery systems experiencing frequent switching between sunlit and eclipse phases, with temperatures fluctuating drastically between -40°C and +60°C (Li, 2024). High-energy particles in space radiation environments cause irreversible capacity loss (Rodrigues et al., 2017), requiring highly autonomous decision-making capabilities (Roman et al., 2021; Zou et al., 2025a). Failure of battery dispatching strategies may interrupt critical missions (Wang et al., 2023).

Current battery management faces severe data scarcity challenges (Schäfer et al., 2024). High-quality operational data remain exceedingly sparse with prohibitively expensive acquisition costs (Tao et al., 2024). While abundant low-quality data contain rich information, they cannot be directly utilised due to sensor noise, inaccurate labelling, and environmental variations, creating a ”data-rich but knowledge-poor” paradox (Alzubaidi et al., 2023). This leads to insufficient model generalisation and decreased reliability (Feng et al., 2020; Severson et al., 2019).

The essence of the problem lies in the fundamental conflict between traditional reductionist modelling paradigms and realistic data characteristics (Wang et al., 2024) as in Fig 1. Traditional methods heavily rely on extensive high-quality annotated data whilst ignoring practical acquisition difficulties (Lanubile et al., 2024). This paradigm assumes identical distributional properties between training and application scenarios, failing to address distribution shifts during long-term satellite



Data cleaning removes low-quality samples by setting quality thresholds, retaining only high-confidence data for training. While ensuring data reliability, it further reduces available samples in data-scarce environments, exacerbating insufficient training data issues.

Data augmentation artificially expands datasets by transforming existing data and adding noise. While increasing data volume, augmented samples are essentially variants of original data, failing to provide genuinely new information and potentially causing model overfitting.

Figure 1: Limitations of traditional data processing methods. Data cleaning removes low-quality samples but exacerbates data scarcity; data augmentation expands data volume but provides no new information and may cause overfitting.

operations (Nozarijouybari & Fathy, 2024). Models exhibit severely inadequate generalisation capabilities when confronted with novel conditions (Lu et al., 2023).

Existing machine learning algorithms exhibit three fundamental deficiencies when applied to data-scarce aerospace energy systems (Attia et al., 2020). Firstly, weighted scalarisation strategies lack theoretical foundation and cannot guarantee Pareto optimality (Peng et al., 2023; Jablonka et al., 2021). Secondly, hard physical constraints are transformed into soft constraints through penalty mechanisms, which cannot strictly guarantee constraint violations will not occur (Pan, 2021; Heenan et al., 2023). Most critically, deep reinforcement learning algorithms heavily depend on data distribution stability assumptions, are prone to local optima, and cannot provide theoretical guarantees for aerospace systems (Wang, 2025; Zou et al., 2025b; Lombardo et al., 2021).

Aerospace energy management systems urgently require a novel paradigm capable of achieving intelligent decision-making in data-scarce environments (Aykol et al., 2020; Lin et al., 2024; Ng et al., 2020). Such systems must incorporate physical constraints as intrinsic components whilst perceiving data quality differences and intelligently utilising mixed-quality information (Beucler et al., 2021; Zhu et al., 2022; Lin et al., 2023; Liu et al., 2025).

We propose an adaptive learning paradigm based on cognitive state awareness that fundamentally redefines the learning problem in data-scarce environments. The core insight lies in recognising that data value is highly dependent on the system's current cognitive level. We construct an exploration-guided learning framework that dynamically adjusts data utilisation strategies by quantifying the system's ”degree of understanding”: when system cognition is limited, conservative learning strategies exploit existing experience; when understanding becomes sufficient, the system transitions towards innovative exploration modes. Through this dynamic matching between cognitive states and learning strategies, we achieve intelligent extraction of valuable information from mixed-quality data.

2 Related Work

2.1 Data Dilemma in Battery Management Systems

Modern battery management systems (BMS) confront a significant data dilemma, as the scarcity of high-quality data and the gap between simulation and reality impede the development of intelligent solutions Zhu et al. (2022); Ng et al. (2020). Data-driven machine learning struggles with a lack of complete lifecycle data due to high acquisition costs and long periods Ling (2022); Aykol et al. (2020). Laboratory data often fails to generalize to real-world conditions Dos Reis et al. (2021); Zou et al. (2023); Lombardo et al. (2021), and the simulation-to-reality gap exacerbates this, as simulations cannot capture complex real-world factors Borah et al. (2024); Liu et al. (2019). Furthermore, online learning is infeasible due to the high risks of thermal runaway Ruan et al. (2021); Feng et al. (2018), and highly variable data quality from real-world sensors introduces noise and missing values, rendering traditional methods ineffective Dubarry et al. (2017); Zhang et al. (2020). These issues collectively pose a fundamental barrier to next-generation intelligent BMS technologies.

2.2 Learning Limitations in Data-Scarce Environments

Existing machine learning (ML) methods face fundamental structural limitations in data-scarce environments, stemming from their static data processing and lack of environmental understanding Liu et al. (2023); Kejriwal et al. (2024). Current approaches use a "one-size-fits-all" strategy that fails to distinguish between high-quality and noisy data, compromising model performance Lombardo et al. (2021); Tuia et al. (2022). Furthermore, these algorithms lack the introspective capability to assess their own understanding of the environment Schulze Buschoff et al. (2025); Voytek & Knight (2015), which prevents them from dynamically adjusting the exploration-exploitation trade-off Yao et al. (2023); Jamieson et al. (2014); Ling (2022). This absence of data value assessment and a reliance on passive learning are key bottlenecks to learning efficiency. Addressing these issues requires a new learning paradigm with cognitive perception that can intelligently utilize mixed-quality data.

3 Problem Definition

Artificial satellites, as mission-critical systems operating autonomously in harsh space environments, face unprecedented technical challenges in coordinated multi-battery management. Satellites experience frequent transitions between sunlit and eclipse phases during 90-minute orbital cycles, with battery temperatures fluctuating dramatically between -40°C and +60°C, whilst simultaneously enduring irreversible capacity losses from radiation. This complex multi-constrained, multi-objective, multi-agent decision-making problem necessitates rigorous mathematical frameworks to guarantee system safety, reliability, and optimality. Consider a satellite power system comprising n battery packs, modelled as a multi-agent Markov decision process $\mathcal{M} = \langle \mathcal{N}, \mathcal{S}, \{\mathcal{A}_i\}, \{\mathcal{T}_i\}, \{\mathcal{R}_i\} \rangle$. The complete state of the i-th battery pack is $s_i(t) = s_i^{local}(t) \cup s^{global}(t)$, where:

$$s_i^{local}(t) = \{SOC_i(t), T_i(t), I_i(t), V_i(t)\} \tag{1}$$

$$s^{global}(t) = \{\theta_{orbit}(t), \phi_{solar}(t), R_{radiation}(t), \mathbf{q}_{attitude}(t), t_{mission}\} \tag{2}$$

The dynamic evolution follows an enhanced equivalent circuit model considering space environment effects:

$$\frac{dSOC_i(t)}{dt} = -\frac{I_i(t)}{Q_{nom,i} \cdot \eta_{temp}(T_i) \cdot \eta_{rad}(R_{radiation})} \tag{3}$$

System operation must satisfy strict constraints including power balance and SOC safety:

$$\sum_{i=1}^{n} P_i(t) = P_{load}(t) + P_{thermal}(t) + P_{attitude}(t) + P_{communication}(t) \tag{4}$$

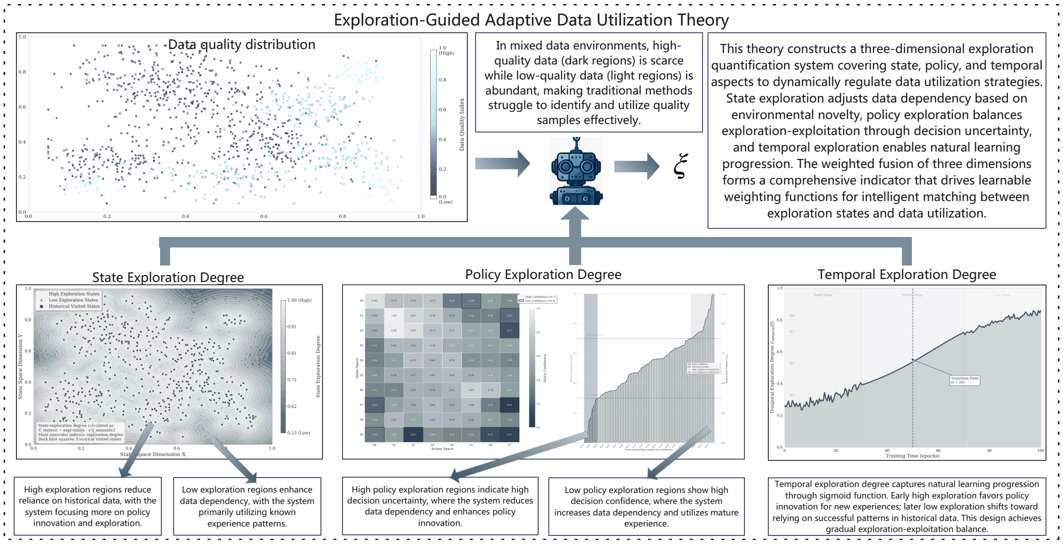
Data acquisition faces significant challenges with mixed-quality dataset $\mathcal{D}_{mixed} = \{(s_j, a_j, r_j, s'_j, q_j)\}_{j=1}^{N}$, where $q_j \sim \text{Beta}(\alpha_{quality}, \beta_{quality})$ denotes implicit data quality labels. Data quality differs due to diverse sources: limited high-fidelity on-orbit data, abundant ground test data with environmental differences, voluminous simulation data with model errors, and poor-quality noisy data. The core challenge is that high-quality data quantity is far smaller than total volume, $|\mathcal{D}_{high}| \ll |\mathcal{D}_{total}|$, whilst quality labels are unknown, rendering traditional methods inapplicable and leading to bias:

$$\mathbb{E}_{\mathcal{D}_{mixed}}[Q_\pi(s,a)] \neq \mathbb{E}_{\rho^\pi}[Q_\pi(s,a)] \tag{5}$$

The control problem is formulated as constrained multi-objective optimisation to learn a policy π minimising:

$$\min_\pi \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \Big(\alpha_1 \bar{C}_{degradation}(s_t, a_t) + \alpha_2 \bar{C}_{energy}(s_t, a_t) - \alpha_3 \bar{R}_{mission}(s_t, a_t) \Big) \right] \tag{6}$$

subject to constraints $g_{power}(\pi) = 0$, $g_{thermal}(\pi) \leq 0$, $g_{SOC}(\pi) \leq 0$, and $\mathbb{P}(\text{constraint violation}) \leq \delta_{safety}$. The core challenges are: learning high-performance policies from mixed-quality data with unknown quality labels, guaranteeing policy generalisation under distribution shifts, and strictly satisfying safety constraints during learning to avoid system failure.


Exploration-Guided Adaptive Data Utilization Theory
Data quality distribution
Data Quality Index
1.0 (High)
0.0 (Low)
ξ
State Exploration Degree
High Exploration States
Low Exploration States
Historical Visited States
State Space Dimension X
State Space Dimension Y
State Exploration Degree
1.00 (High)
0.53 (Low)
State exploration degree calculated as:
E_state(s) = exp(-min||s - s'||_semantic)
Point size/color indicate exploration degree
Dark blue squares: historical visited states
Policy Exploration Degree
High Confidence (>0.7)
Low Confidence (<0.3)
Action Space
State Space
Policy Confidence
State-Action Pairs (sorted by confidence)
Temporal Exploration Degree
Early Stage
Middle Stage
Late Stage
Transition Point (t = 50)
Temporal Exploration Degree $\mathcal{E}_{temporal}(t)$
Training Time (epochs)

In mixed data environments, high-quality data (dark regions) is scarce while low-quality data (light regions) is abundant, making traditional methods struggle to identify and utilize quality samples effectively.

This theory constructs a three-dimensional exploration quantification system covering state, policy, and temporal aspects to dynamically regulate data utilization strategies. State exploration adjusts data dependency based on environmental novelty, policy exploration balances exploration-exploitation through decision uncertainty, and temporal exploration enables natural learning progression. The weighted fusion of three dimensions forms a comprehensive indicator that drives learnable weighting functions for intelligent matching between exploration states and data utilization.

High exploration regions reduce reliance on historical data, with the system focusing more on policy innovation and exploration.

Low exploration regions enhance data dependency, with the system primarily utilizing known experience patterns.

High policy exploration regions indicate high decision uncertainty, where the system reduces data dependency and enhances policy innovation.

Low policy exploration regions show high decision confidence, where the system increases data dependency and utilizes mature experience.

Temporal exploration degree captures natural learning progression through sigmoid function. Early high exploration favors policy innovation for new experiences; later low exploration shifts toward relying on successful patterns in historical data. This design achieves gradual exploration-exploitation balance.

Figure 2: Exploration-guided adaptive data utilization theory. The theory constructs a three-dimensional exploration quantification system (state, policy, and temporal exploration degrees) to intelligently extract valuable information from mixed-quality data through dynamic matching between exploration states and data utilization strategies.

4 Methods

4.1 Exploration-Guided Adaptive Data Utilisation Theory

When confronting mixed-quality data environments, the core limitation of traditional reinforcement learning methods lies in their inability to intelligently identify and utilise the potential value of different data. As in Fig 2, we propose an exploration-guided adaptive data utilisation theory that dynamically adjusts data dependency strategies by quantifying the system's "degree of understanding" of the environment. The core insight is that when the system has limited understanding of the environment, it should rely more heavily on experiential patterns in historical data; whilst when the system possesses sufficient understanding, it can more freely engage in policy innovation.

The quantification of exploration degree constitutes the foundation of this theoretical framework. We define exploration degree as a composite indicator of state novelty, policy uncertainty, and learning progress:

$$\mathcal{E}(s, \pi, t) = \alpha_1 \mathcal{E}_{state}(s) + \alpha_2 \mathcal{E}_{policy}(\pi) + \alpha_3 \mathcal{E}_{temporal}(t) \tag{7}$$

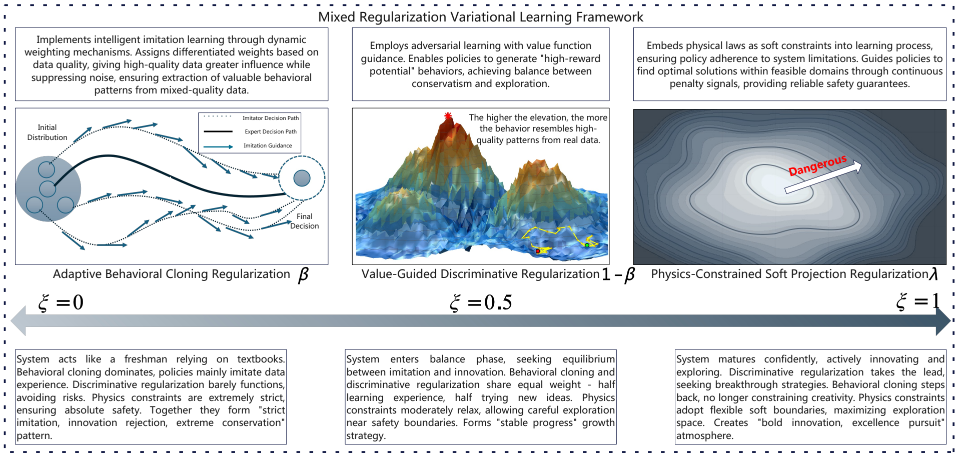
where state exploration degree measures the novelty of the current state, policy exploration degree reflects uncertainty and diversity of policy outputs, and temporal exploration degree captures the stage characteristics of the learning process. The dynamic evolution of exploration degree follows:

$$\frac{d\mathcal{E}}{dt} = -\gamma_{\text{decay}}\mathcal{E} + \beta_{\text{surprise}}\mathcal{I}_{\text{surprise}}(s_t, a_t, r_t) - \lambda_{\text{constraint}}\mathcal{V}_{\text{violation}}(s_t, a_t) \tag{8}$$

Exploration degree-based data weight learning constitutes the core innovation of this theory. Unlike traditional methods that pre-evaluate data quality, we design a learnable weight function:

$$w_{\text{adaptive}}(s, a, r, s') = \sigma\left(\mathbf{W}_w^T \phi_{\text{context}}(s, a, r, s', \mathcal{E})\right) \tag{9}$$

The theoretical advantage lies in achieving end-to-end joint optimisation of weight learning and policy learning. The weight function automatically learns to identify which data samples possess higher value at the current learning stage during training, without relying on pre-designed heuristic rules. This exploration degree-guided adaptive mechanism provides a principled approach for handling mixed data environments with unknown data quality, avoiding the difficulties of pre-evaluating data quality.


Mixed Regularization Variational Learning Framework
Initial Distribution
Imitator Decision Path
Expert Decision Path
Imitation Guidance
Final Decision
Dangerous
Adaptive Behavioral Cloning Regularization β
Value-Guided Discriminative Regularization $1-\beta$
Physics-Constrained Soft Projection Regularization λ
$\xi=0$
$\xi=0.5$
$\xi=1$

Implements intelligent imitation learning through dynamic weighting mechanisms. Assigns differentiated weights based on data quality, giving high-quality data greater influence while suppressing noise, ensuring extraction of valuable behavioral patterns from mixed-quality data.

Employs adversarial learning with value function guidance. Enables policies to generate "high-reward potential" behaviors, achieving balance between conservatism and exploration.

Embeds physical laws as soft constraints into learning process, ensuring policy adherence to system limitations. Guides policies to find optimal solutions within feasible domains through continuous penalty signals, providing reliable safety guarantees.

The higher the elevation, the more the behavior resembles high-quality patterns from real data.

System acts like a freshman relying on textbooks. Behavioral cloning dominates, policies mainly imitate data experience. Discriminative regularization barely functions, avoiding risks. Physics constraints are extremely strict, ensuring absolute safety. Together they form "strict imitation, innovation rejection, extreme conservation" pattern.

System enters balance phase, seeking equilibrium between imitation and innovation. Behavioral cloning and discriminative regularization share equal weight - half learning experience, half trying new ideas. Physics constraints moderately relax, allowing careful exploration near safety boundaries. Forms "stable progress" growth strategy.

System matures confidently, actively innovating and exploring. Discriminative regularization takes the lead, seeking breakthrough strategies. Behavioral cloning steps back, no longer constraining creativity. Physics constraints adopt flexible soft boundaries, maximizing exploration space. Creates "bold innovation, excellence pursuit" atmosphere.

Figure 3: Mixed regularization variational learning framework. The framework integrates adaptive behavioral cloning, value-guided discriminator, and physics-constrained regularization mechanisms, dynamically adjusting regularization weights through exploration degree to achieve intelligent balance between conservatism and exploration.

4.2 Mixed Regularization Variational Learning Framework

Building upon the exploration-guided adaptive data utilisation foundation, we construct a mixed regularization variational learning framework to unifiedly address multiple learning objectives including behavioural cloning, exploration guidance, and physical constraint satisfaction as in Fig 3. Traditional offline reinforcement learning methods often rely on single regularization mechanisms, struggling to find optimal balance between conservatism and exploration. Our framework organically integrates multiple regularization strategies through variational inference theory.

The variational learning framework reformulates the policy optimisation problem as a variational inference problem. We seek the optimal policy that maximises:

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{(s,a)\sim\mathcal{D}}[w(s,a)\log\pi(a|s)] - \mathrm{KL}(\pi\|p(\pi)) \tag{10}$$

The mixed regularization mechanism organically combines three different types of regularization terms:

$$\mathcal{R}_{\text{blended}}(\pi) = (1-\beta)\mathcal{R}_{\text{BC}}(\pi) + \beta\mathcal{R}_{\text{DR}}(\pi) + \lambda_{\text{phys}}\mathcal{R}_{\text{physics}}(\pi) \tag{11}$$

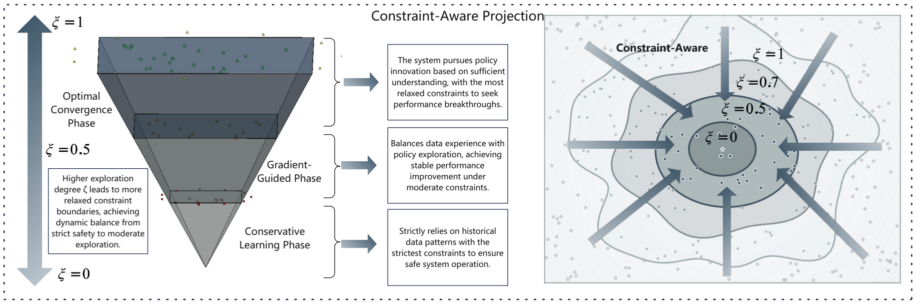
where the behavioural cloning regularization term ensures policy consistency with good behavioural patterns, the discriminator regularization term guides policy exploration towards high-value regions through adversarial learning mechanisms, and the physics constraint regularization term ensures learned policies obey the physical laws of satellite systems. Adaptive adjustment of balance parameters constitutes the key mechanism for achieving exploration-guided learning:

$$\beta(\mathcal{E}) = \min(1, \beta_0 + \beta_{\text{adaptive}} \cdot \mathcal{E}) \tag{12}$$

This design ensures that the learning process can automatically adjust the balance between conservatism and exploration according to the system's current understanding level. The theoretical advantages manifest in providing a principled mathematical framework for multi-objective fusion, achieving automatic identification and utilisation of data quality, and ensuring smooth transitions from data-driven to model-driven approaches.

4.3 Exploration-Aware Actor-Critic Learning Algorithm

Based on the exploration degree theory and mixed regularization variational learning framework, we design an exploration-aware Actor-Critic learning algorithm that deeply integrates exploration degree mechanisms into the core components of policy gradient and value function estimation. Traditional Actor-Critic methods often exhibit instability when handling mixed-quality data due to their lack of awareness regarding data quality and learning states in gradient estimation and value propagation processes.



Higher exploration degree ζ leads to more relaxed constraint boundaries, achieving dynamic balance from strict safety to moderate exploration.

The system pursues policy innovation based on sufficient understanding, with the most relaxed constraints to seek performance breakthroughs.

Balances data experience with policy exploration, achieving stable performance improvement under moderate constraints.

Strictly relies on historical data patterns with the strictest constraints to ensure safe system operation.

Figure 4: Constraint-aware projection and satellite environment specialization theory. The theory dynamically modulates constraint strictness through exploration degree, progressively transitioning from strict constraints in conservative learning phase to moderate constraint relaxation in optimal convergence phase, achieving dynamic balance between safety and performance.

Exploration-aware modification of policy gradients constitutes one of the core innovations. We dynamically adjust weights according to exploration degree and sample characteristics:

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{(s,a)\sim\rho^\pi}[w_{\mathcal{E}}(s,a) \cdot Q^\pi(s,a) \cdot \nabla_\theta \log \pi_\theta(a|s)] \tag{13}$$

The dual-scale design of value function estimation specifically targets the temporal characteristics of satellite multi-battery systems, separately modeling rapid environmental changes within orbital cycles and slow-varying processes across orbital cycles:

$$Q_{\text{dual}}(s,a) = \alpha_{\text{short}} Q_{\text{short}}(s,a) + \alpha_{\text{long}} Q_{\text{long}}(s,a) \tag{14}$$

The exploration degree-modulated Bellman equation embeds exploration-aware mechanisms into the fundamental update rules of value functions:

$$Q^\pi(s,a) = r(s,a) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a), a'\sim\pi(\cdot|s')}[w_{\mathcal{E}}(s',a') \cdot Q^\pi(s',a')] \tag{15}$$

Joint training of Actor and Critic networks follows exploration-aware update rules that integrate mixed regularization variational objectives with exploration-aware policy gradients. The practical significance lies in providing a complete solution that can both fully utilize mixed-quality data and guarantee learning stability.

4.4 Constraint-Aware Projection and Satellite Environment Specialization Theory

Building upon the exploration-aware Actor-Critic algorithm, we establish constraint-aware projection operator theory and satellite environment specialization theory extensions to ensure algorithmic safety under complex constraint conditions and robustness in extreme space environments as in Fig 4. These theoretical modules jointly constitute a complete theoretical framework for satellite multi-battery management.

The design of constraint-aware projection operators constitutes the key to guaranteeing safe system operation. The feasible policy space for satellite multi-battery systems is defined as:

$$\Pi_{satellite} = \bigcap_{i=1}^{m} \Pi_i \tag{16}$$

We propose exploration degree-modulated adaptive projection mechanisms:

$$P_{\mathcal{E}}[x] = (1-\mathcal{E})P_{strict}[x] + \mathcal{E}P_{relaxed}[x] \tag{17}$$

When exploration degree is low, the system employs strict projection ensuring policies strictly satisfy all constraints; as exploration degree increases, relaxed projection allows limited exploration near constraint boundaries. Satellite environment specialization theory extensions are manifested in orbital periodicity modeling, extreme environment robustness, and distributed learning consistency. The 90-minute periodicity of satellite operations can be represented using Fourier series, enabling algorithms to naturally capture and utilize periodic patterns. Robustness measures ensure policies maintain acceptable performance under worst-case scenarios including drastic temperature changes, radiation intensity fluctuations, and attitude control errors. Distributed learning consistency addresses multi-satellite coordination scenarios, guaranteeing convergence to globally consistent solutions under connected communication topology conditions.

4.5 Theoretical Properties and Convergence Analysis

Based on the theoretical framework, we conduct comprehensive analysis of the theoretical properties encompassing sample complexity, constraint satisfaction guarantees, and asymptotic optimality. Sample complexity analysis reveals that under standard assumptions, the algorithm's sample complexity is:

$$\mathcal{O}\left(\frac{|\mathcal{S}||\mathcal{A}|\log(1/\delta)}{(1-\gamma)^4\epsilon^2}\cdot\left(1+\frac{\mathrm{Var}[\mathcal{E}]}{\mathbb{E}[\mathcal{E}]^2}\right)\right) \tag{18}$$

Probabilistic guarantees of constraint satisfaction possess a lower bound:

$$\mathbb{P}(\text{constraint satisfaction}) \geq 1-\delta_{safety}\cdot\exp(-\kappa(1-\mathcal{E})) \tag{19}$$

When exploration degree is low, the system automatically provides higher safety guarantees; as exploration degree increases, constraint satisfaction probability decreases somewhat but remains within acceptable ranges. Asymptotic optimality guarantees that exploration degree-guided policies asymptotically converge to optimal policies under constraint conditions:

$$\lim_{N\to\infty} J(\pi_N) = J(\pi^*) \tag{20}$$

Comprehensively, this theoretical framework provides complete mathematical foundations for learning satellite multi-battery management policies from mixed-quality data. Exploration degree-guided mechanisms address the core challenge of unknown data quality whilst providing adaptive adjustment capabilities. Constraint-aware projection and environment specialization extensions ensure algorithmic safety and robustness in actual deployment.

5 Experiments

5.1 Experimental Setup and Simulation Platform

To validate the effectiveness of the proposed exploration degree-guided learning framework in satellite multi-battery systems, we constructed a high-fidelity satellite simulation platform. This platform integrates orbital dynamics models, thermodynamic models, electrochemical models, and radiation environment models, enabling realistic simulation of satellite operational states in space environments. The simulation considers the actual environmental characteristics of satellite orbital operations with 90-minute orbital periods, requiring the battery system to frequently switch operating modes between sunlit and eclipse phases, with temperatures fluctuating dramatically between -40°C and +60°C. Power requirements vary with orbital position: in sunlit regions, battery burden is relatively light due to solar panel power supply, whilst eclipse regions rely entirely on battery power. Considering the difficulties of data acquisition and quality variations in real satellite operations, we constructed four datasets of different quality levels with expert data ratios ranging from 0% to 100%, each containing 400k time-step transition samples covering approximately 280 complete orbital cycles.

All metrics are calculated based on 3600 seconds (4 complete orbital cycles) of simulation time, with experiments repeated 10 times to obtain average values and standard deviations. Power tracking accuracy evaluates the system's ability to satisfy power requirements:

$$\eta_{tracking} = \left(1-\sqrt{\frac{1}{T}\int_0^T\left(\frac{P_{output}(t)-P_{demand}(t)}{P_{demand}(t)}\right)^2 dt}\right)\times 100\% \tag{21}$$

Energy efficiency is measured through the ratio of useful energy output to total consumed energy:

$$\eta_{energy} = \frac{\int_0^T P_{useful}(t)dt}{\int_0^T P_{total}(t)dt}\times 100\% \tag{22}$$

Data quality robustness evaluation measures algorithm sensitivity to data quality variations through coefficient of variation:

$$CV = \frac{\sigma}{\mu}\times 100\% \tag{23}$$

where σ represents the standard deviation of a performance metric across four different quality datasets, and μ represents the mean of that metric across the four datasets. Smaller coefficient of variation indicates lower algorithm sensitivity to data quality changes and stronger robustness.

5.2 Experimental Results

The experimental results demonstrate the comprehensive performance of our method across multiple evaluation metrics. Table 1 present detailed comparisons of all algorithms across four datasets for power tracking accuracy and energy efficiency performance.

Table 1: Comprehensive Performance Comparison Across Core Metrics

Power Tracking Accuracy (%)

Algorithm	D1 Dataset	D2 Dataset	D3 Dataset	D4 Dataset	Average Accuracy
Our Method	**90.7±1.4**	**89.2±1.8**	**85.3±2.6**	**76.8±3.8**	**85.5±2.4**
PPO	91.2±1.3	87.8±2.1	81.4±3.2	72.5±4.3	83.2±2.7
TD3	90.9±1.5	87.4±2.4	80.9±3.4	71.8±4.6	82.8±3.0
SAC	91.0±1.4	88.1±2.2	81.7±3.1	73.2±4.1	83.5±2.7
BC	89.8±1.7	84.3±2.5	75.2±3.8	Convergence Failed	83.1±2.7
CQL	90.3±1.6	87.6±2.3	82.9±2.9	74.6±3.4	83.9±2.6

Energy Efficiency (%)

Algorithm	D1 Dataset	D2 Dataset	D3 Dataset	D4 Dataset	Average Efficiency
Our Method	82.4±1.7	**81.8±2.1**	**79.2±2.8**	**75.1±3.6**	**79.6±2.6**
PPO	**83.1±1.5**	80.9±2.2	76.8±3.4	71.2±4.2	78.0±2.8
TD3	82.8±1.8	80.4±2.4	76.1±3.6	70.8±4.5	77.5±3.1
SAC	83.0±1.6	81.2±2.3	77.3±3.2	72.6±4.0	78.5±2.8
BC	81.7±2.0	79.3±2.7	72.8±4.1	Convergence Failed	77.9±2.9
CQL	82.1±1.9	80.8±2.5	78.1±3.0	74.3±3.8	78.8±2.8

Power tracking performance directly reflects the system's ability to satisfy satellite load requirements. Experimental results demonstrate that on high-quality dataset D1, traditional online reinforcement learning methods perform slightly better than our method, which meets expectations since sufficient exploration under ideal data conditions can yield superior performance. However, as data quality deteriorates, our method's exploration degree-guided mechanism begins to take effect. On datasets D2 and D3, our method achieves tracking accuracies of 89.2% and 85.3% respectively, improving over the best baseline methods by approximately 1.4% and 3.6%. On the extreme D4 dataset, our method still maintains 76.8% tracking accuracy, representing approximately 4.3% improvement over traditional methods. In terms of energy efficiency, our method maintains relatively stable performance as data quality decreases. On datasets D3 and D4, our method achieves efficiencies of 79.2% and 75.1% respectively, improving over the best baseline methods by approximately 1.9% and 2.5%. This progressive performance advantage reflects the effectiveness of exploration degree-guided mechanisms in handling data uncertainty.

Table 2: Performance Metric Coefficient of Variation Comparison (%)

Algorithm	Power Tracking Accuracy	Energy Efficiency
Our Method	**7.8**	**4.6**
PPO	9.2	6.1
TD3	9.5	6.4
SAC	8.7	5.8
BC	8.9	5.9
CQL	8.3	5.2

To evaluate algorithm sensitivity to data quality variations, we analysed the degree of variation in performance metrics across different datasets as shown in Table 2. Coefficient of variation analysis demonstrates that our method exhibits the best data quality robustness across all performance metrics. The coefficient of variation for power tracking accuracy is 7.8%, representing approximately 0.9% improvement over the best baseline method. The coefficient of variation for energy efficiency is 4.6%, representing approximately 0.6% improvement over the best baseline method. Whilst these improvement margins are moderate, they possess important practical value in aerospace applications.

6 Discussion

Experimental results demonstrate that our proposed exploration degree-guided adaptive learning framework exhibits encouraging performance in satellite multi-battery management tasks, particularly demonstrating robustness advantages under mixed-quality data environments. The framework achieves performance improvements of 3.6% and 2.5% in power tracking accuracy and energy efficiency metrics respectively on challenging datasets. More importantly, the coefficients of variation for performance metrics are significantly reduced, proving the algorithm's strong adaptability to data quality variations.

The fundamental reason for our framework's effectiveness lies in the cognitive transformation of traditional reinforcement learning paradigms. Traditional methods treat all data equally, whilst our exploration degree mechanism achieves intelligent identification and dynamic utilization of data value. When system understanding of the environment is limited, the exploration degree-guided mechanism prompts the algorithm to rely more heavily on reliable patterns in historical experience; as system understanding deepens, increasing exploration degree gradually weakens dependence on historical data, instead guiding policy exploration through discriminator regularization. This adaptive mechanism avoids the limitations of fixed weight design, enabling learning processes to intelligently balance conservatism and exploration according to current cognitive states.

The profound significance of this adaptive mechanism lies in its breakthrough contribution to decision-making problems under incomplete information environments. Our exploration degree-guided framework fundamentally redefines this challenge by quantifying system cognitive uncertainty, transforming information deficiency from an obstacle into an intelligently manageable system characteristic. This paradigmatic shift establishes a dynamic bridge between system ”self-knowledge” and ”decision-making”—the more systems understand their cognitive limitations, the wiser decisions they can make. From an industrial application perspective, this framework directly addresses core technical challenges in aerospace engineering, providing direct application value for missions including unmanned deep space exploration, commercial satellite constellation management, and long-term space station operations.

Future research will focus on constructing more intelligent multi-objective balancing frameworks, reformulating the regularization terms as multi-objective optimization problems whilst strictly modeling physical constraints as hard constraint conditions. Additionally, further refinement of exploration degree quantification systems, particularly extensions in multi-time-scale and multi-agent environments, will establish more solid theoretical foundations for constructing intelligent systems with genuine cognitive adaptive capabilities.

7 Conclusion

This research addresses the intelligent management challenges of satellite multi-battery systems in data-scarce environments, innovatively proposing an exploration degree-guided adaptive learning framework. This framework achieves the goal of intelligently extracting valuable information from mixed-quality data by quantifying the system's ”degree of understanding” of the environment. Core contributions include: constructing an exploration degree measurement system that unifiedly guides data utilization, constraint adjustment, and exploration decisions; designing a mixed regularization variational learning framework that integrates behavioural cloning, discriminator regularization, and physics constraint regularization; developing an exploration-aware Actor-Critic algorithm that deeply integrates exploration degree mechanisms into policy learning components. Experimental results demonstrate that the proposed method significantly outperforms baseline methods in satellite multi-battery management tasks.

References

Laith Alzubaidi, Jinshuai Bai, Aiman Al-Sabaawi, Jose Santamaría, Ahmed Shihab Albahri, Bashar Sami Nayyef Al-Dabbagh, Mohammed A Fadhel, Mohamed Manoufali, Jinglan Zhang, Ali H Al-

Timemy, et al. A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. *Journal of Big Data*, 10(1):46, 2023.

Peter M Attia, Aditya Grover, Norman Jin, Kristen A Severson, Todor M Markov, Yang-Hung Liao, Michael H Chen, Bryan Cheong, Nicholas Perkins, Zi Yang, et al. Closed-loop optimization of fast-charging protocols for batteries with machine learning. *Nature*, 578(7795):397–402, 2020.

Muratahan Aykol, Patrick Herring, and Abraham Anapolsky. Machine learning for continuous innovation in battery technologies. *Nature Reviews Materials*, 5(10):725–727, 2020.

Tom Beucler, Michael Pritchard, Stephan Rasp, Jordan Ott, Pierre Baldi, and Pierre Gentine. Enforcing analytic constraints in neural networks emulating physical systems. *Physical review letters*, 126(9):098302, 2021.

Manashita Borah, Qiao Wang, Scott Moura, Dirk Uwe Sauer, and Weihan Li. Synergizing physics and machine learning for advanced battery management. *Communications Engineering*, 3(1):134, 2024.

Gonçalo Dos Reis, Calum Strange, Mohit Yadav, and Shawn Li. Lithium-ion battery data and where to find it. *Energy and AI*, 5:100081, 2021.

Matthieu Dubarry, Arnaud Devie, and Katherine McKenzie. Durability and reliability of electric vehicle batteries under electric utility grid operations: Bidirectional charging impact analysis. *Journal of Power Sources*, 358:39–49, 2017.

Xuning Feng, Minggao Ouyang, Xiang Liu, Languang Lu, Yong Xia, and Xiangming He. Thermal runaway mechanism of lithium ion battery for electric vehicles: A review. *Energy storage materials*, 10:246–267, 2018.

Xuning Feng, Dongsheng Ren, Xiangming He, and Minggao Ouyang. Mitigating thermal runaway of lithium-ion batteries. *Joule*, 4(4):743–770, 2020.

TMM Heenan, I Mombrini, A Llewellyn, S Checchia, C Tan, MJ Johnson, Anmol Jnawali, Gaston Garbarino, Rhodri Jervis, DJL Brett, et al. Mapping internal temperatures during high-rate battery applications. *Nature*, 617(7961):507–512, 2023.

Kevin Maik Jablonka, Giriprasad Melpatti Jothiappan, Shefang Wang, Berend Smit, and Brian Yoo. Bias free multiobjective active learning for materials design and discovery. *Nature communications*, 12(1):2312, 2021.

Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil'ucb: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, pp. 423–439. PMLR, 2014.

Mayank Kejriwal, Eric Kildebeck, Robert Steininger, and Abhinav Shrivastava. Challenges, evaluation and opportunities for open-world learning. *Nature Machine Intelligence*, 6(6):580–588, 2024.

Andrea Lanubile, Pietro Bosoni, Gabriele Pozzato, Anirudh Allam, Matteo Acquarone, and Simona Onori. Domain knowledge-guided machine learning framework for state of health estimation in lithium-ion batteries. *Communications Engineering*, 3(1):168, 2024.

Ge Li. Promotion of practical technology of the thermal management system for cylindrical power battery. *Energy Informatics*, 7(1):33, 2024.

Zihui Lin, Dagang Li, and Yuntao Zou. Energy efficiency of lithium-ion batteries: Influential factors and long-term degradation. *Journal of Energy Storage*, 74:109386, 2023.

Zihui Lin, Zhongwei Huang, Shuojin Yang, Chenhao Wu, Sen Fang, Zhichun Liu, Min Yang, Zhihui Xie, Dagang Li, and Yuntao Zou. Survey on task-centric robot battery management: A neural network framework. *Journal of Power Sources*, 610:234674, 2024.

Chen Ling. A review of the recent progress in battery informatics. *npj Computational Materials*, 8 (1):33, 2022.

Bo Liu, Xintong Yuan, and Yuzhang Li. Colossal capacity loss during calendar aging of zn battery chemistries. *ACS Energy Letters*, 8(9):3820–3828, 2023.

Hongao Liu, Chang Li, Xiaosong Hu, Jinwen Li, Kai Zhang, Yang Xie, Ranglei Wu, and Ziyou Song. Multi-modal framework for battery state of health evaluation using open-source electric vehicle data. *Nature Communications*, 16(1):1137, 2025.

Kailong Liu, Yi Li, Xiaosong Hu, Mattin Lucu, and Widanalage Dhammika Widanage. Gaussian process regression with automatic relevance determination kernel for calendar aging prediction of lithium-ion batteries. *IEEE Transactions on Industrial Informatics*, 16(6):3767–3777, 2019.

Teo Lombardo, Marc Duquesnoy, Hassna El-Bouysidy, Fabian Årén, Alfonso Gallo-Bueno, Peter Bjørn Jørgensen, Arghya Bhowmik, Arnaud Demortiere, Elixabete Ayerbe, Francisco Alcaide, et al. Artificial intelligence applied to battery research: hype or reality? *Chemical reviews*, 122 (12):10899–10969, 2021.

Jiahuan Lu, Rui Xiong, Jinpeng Tian, Chenxu Wang, and Fengchun Sun. Deep learning to estimate lithium-ion battery state of health without additional degradation experiments. *Nature Communications*, 14(1):2760, 2023.

Man-Fai Ng, Jin Zhao, Qingyu Yan, Gareth J Conduit, and Zhi Wei Seh. Predicting the state of charge and health of batteries using data-driven machine learning. *Nature Machine Intelligence*, 2(3):161–170, 2020.

Zahra Nozarijouybari and Hosam K Fathy. Machine learning for battery systems applications: Progress, challenges, and opportunities. *Journal of Power Sources*, 601:234272, 2024.

Jie Pan. Can machines learn with hard constraints? *Nature Computational Science*, 1(4):244–244, 2021.

Bo Peng, Ye Wei, Yu Qin, Jiabao Dai, Yue Li, Aobo Liu, Yun Tian, Liuliu Han, Yufeng Zheng, and Peng Wen. Machine learning-enabled constrained multi-objective design of architected materials. *Nature Communications*, 14(1):6630, 2023.

Zhonghao Rao, Peizhao Lyu, Menghan Li, Xinjian Liu, and Xuning Feng. A thermal perspective on battery safety. *Nature Reviews Clean Technology*, pp. 1–14, 2025.

Marco-Tulio F Rodrigues, Ganguli Babu, Hemtej Gullapalli, Kaushik Kalaga, Farheen N Sayed, Keiko Kato, Jarin Joyner, and Pulickel M Ajayan. A materials perspective on li-ion batteries at extreme temperatures. *nature energy*, 2(8):1–14, 2017.

Darius Roman, Saurabh Saxena, Valentin Robu, Michael Pecht, and David Flynn. Machine learning pipeline for battery state-of-health estimation. *Nature Machine Intelligence*, 3(5):447–456, 2021.

Haijun Ruan, Bingxiang Sun, Jiuchun Jiang, Weige Zhang, Xitian He, Xiaojia Su, Jingji Bian, and Wenzhong Gao. A modified-electrochemical impedance spectroscopy-based multi-time-scale fractional-order model for lithium-ion batteries. *Electrochimica Acta*, 394:139066, 2021.

Raphael Schäfer, Till Nicke, Henning Höfener, Annkristin Lange, Dorit Merhof, Friedrich Feuerhake, Volkmar Schulz, Johannes Lotz, and Fabian Kiessling. Overcoming data scarcity in biomedical imaging with a foundational multi-task model. *Nature Computational Science*, 4(7):495–509, 2024.

Luca M Schulze Buschoff, Elif Akata, Matthias Bethge, and Eric Schulz. Visual cognition in multimodal large language models. *Nature Machine Intelligence*, 7(1):96–106, 2025.

Kristen A Severson, Peter M Attia, Norman Jin, Nicholas Perkins, Benben Jiang, Zi Yang, Michael H Chen, Muratahan Aykol, Patrick K Herring, Dimitrios Fraggedakis, et al. Data-driven prediction of battery cycle life before capacity degradation. *Nature Energy*, 4(5):383–391, 2019.

Shengyu Tao, Ruifei Ma, Zixi Zhao, Guangyuan Ma, Lin Su, Heng Chang, Yuou Chen, Haizhou Liu, Zheng Liang, Tingwei Cao, et al. Generative learning assisted state-of-health estimation for sustainable battery recycling with random retirement conditions. *Nature Communications*, 15(1): 10154, 2024.

Devis Tuia, Benjamin Kellenberger, Sara Beery, Blair R Costelloe, Silvia Zuffi, Benjamin Risse, Alexander Mathis, Mackenzie W Mathis, Frank Van Langevelde, Tilo Burghardt, et al. Perspectives in machine learning for wildlife conservation. *Nature communications*, 13(1):792, 2022.

Bradley Voytek and Robert T Knight. Dynamic network communication as a unifying neural basis for cognition, development, aging, and disease. *Biological psychiatry*, 77(12):1089–1097, 2015.

Fujin Wang, Zhi Zhai, Zhibin Zhao, Yi Di, and Xuefeng Chen. Physics-informed neural network for lithium-ion battery degradation stable modeling and prognosis. *Nature Communications*, 15 (1):4332, 2024.

Ying Wang. Application-oriented design of machine learning paradigms for battery science. *npj Computational Materials*, 11(1):89, 2025.

Yujie Wang, Xingchen Zhang, Kaiquan Li, Guanghui Zhao, and Zonghai Chen. Perspectives and challenges for future lithium-ion battery control and management, 2023.

Zhenpeng Yao, Yanwei Lum, Andrew Johnston, Luis Martin Mejia-Mendoza, Xin Zhou, Yonggang Wen, Alán Aspuru-Guzik, Edward H Sargent, and Zhi Wei Seh. Machine learning for a sustainable energy future. *Nature Reviews Materials*, 8(3):202–215, 2023.

Yunwei Zhang, Qiaochu Tang, Yao Zhang, Jiabin Wang, Ulrich Stimming, and Alpha A Lee. Identifying degradation patterns of lithium ion batteries from impedance spectroscopy using machine learning. *Nature communications*, 11(1):1706, 2020.

Jiangong Zhu, Yixiu Wang, Yuan Huang, R Bhushan Gopaluni, Yankai Cao, Michael Heere, Martin J Mühlbauer, Liuda Mereacre, Haifeng Dai, Xinhua Liu, et al. Data-driven capacity estimation of commercial lithium-ion batteries from voltage relaxation. *Nature communications*, 13(1):2261, 2022.

Yuntao Zou, Zihui Lin, Dagang Li, and ZhiChun Liu. Advancements in artificial neural networks for health management of energy storage lithium-ion batteries: A comprehensive review. *Journal of Energy Storage*, 73:109069, 2023.

Yuntao Zou, Zeling Xu, Tingting Wang, Gang Xiong, Zihui Lin, and Dagang Li. Generative ai-driven dynamic information prioritization for enhanced autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 2025a.

Yuntao Zou, Zeling Xu, Qianqi Zhang, Zihui Lin, Tingting Wang, Zhichun Liu, and Dagang Li. Few-shot learning with manifold-enhanced llm for handling anomalous perception inputs in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 2025b.