Adaptive Pareto-Optimal and Generalizable Multi-Objective Offloading and Resource Scheduling in Dynamic Mobile Edge Computing for Enhanced User Experience

Xuan Li * and Haoran Zuo

Posted Date: 26 March 2026

doi: 10.20944/preprints202603.2141.v1

Keywords: mobile edge computing; multi-agent reinforcement learning; multi-objective optimization; resource management; adaptation

 doi:10.20944/preprints202603.2141.v1

Article

Adaptive Pareto-Optimal and Generalizable Multi-Objective Offloading and Resource Scheduling in Dynamic Mobile Edge Computing for Enhanced User Experience

Xuan Li and Haoran Zuo

Henan Polytechnic University; 202138590215@stu.kust.edu.cn

Abstract

Mobile Edge Computing (MEC) faces significant challenges in dynamic environments, balancing conflicting objectives like latency, energy, and Quality of Experience (QoE) amidst heterogeneous resources and multi-user competition. These issues are compounded by poor generalization and slow adaptation. This paper introduces APOG-MARL (Adaptive Pareto-Optimal and Generalizable Multi-Agent Reinforcement Learning), a novel framework built on an Adaptive-Contextual Multi-Objective Markov Decision Process (MOMDP). APOG-MARL integrates a hierarchical context-aware state representation for generalization, a multi-objective Pareto policy network for optimal trade-offs, a constraint-driven multi-agent collaboration mechanism for efficient resource management, and a meta-learning approach for rapid user preference adaptation. Extensive simulations demonstrate APOG-MARL's superior performance across varying network scales, dynamic user preferences, and high resource utilization scenarios. It achieves enhanced user QoE, significantly lower average task latency and total energy consumption, superior Pareto front quality, and robust resource utilization, consistently outperforming state-of-the-art baselines. APOG-MARL offers a powerful and practical solution for optimizing task offloading and resource scheduling in complex, dynamic MEC environments.

Keywords: mobile edge computing; multi-agent reinforcement learning; multi-objective optimization; resource management; adaptation

1. Introduction

Mobile Edge Computing (MEC) has emerged as a pivotal paradigm, extending computational and storage resources closer to end-users at the network edge [1]. This proximity significantly reduces task latency and alleviates the load on core networks, providing robust support for latency-sensitive and high-bandwidth applications such as the Internet of Things (IoT), 5G/6G communications, and Augmented Reality (AR) [2]. The ability of MEC to enhance responsiveness and improve user experience underscores its critical importance in modern distributed computing landscapes.

Despite its immense potential, the dynamic and heterogeneous nature of MEC environments, coupled with intense multi-user competition for limited resources, presents formidable challenges for effective task offloading and resource scheduling [3]. Traditional approaches often suffer from several key limitations:

- **Single-Objective Optimization Dilemma**: Most existing works predominantly focus on optimizing a single metric, such as minimizing latency or energy consumption. This narrow focus struggles to achieve a Pareto-optimal trade-off among multiple conflicting objectives, including latency, energy, and Quality of Experience (QoE), which are crucial for real-world application scenarios [4].
- **Insufficient Generalization Capability**: Current methods are typically trained and evaluated in fixed-scenario setups. Their performance degrades sharply when environmental parameters, such

 doi:10.20944/preprints202603.2141.v1

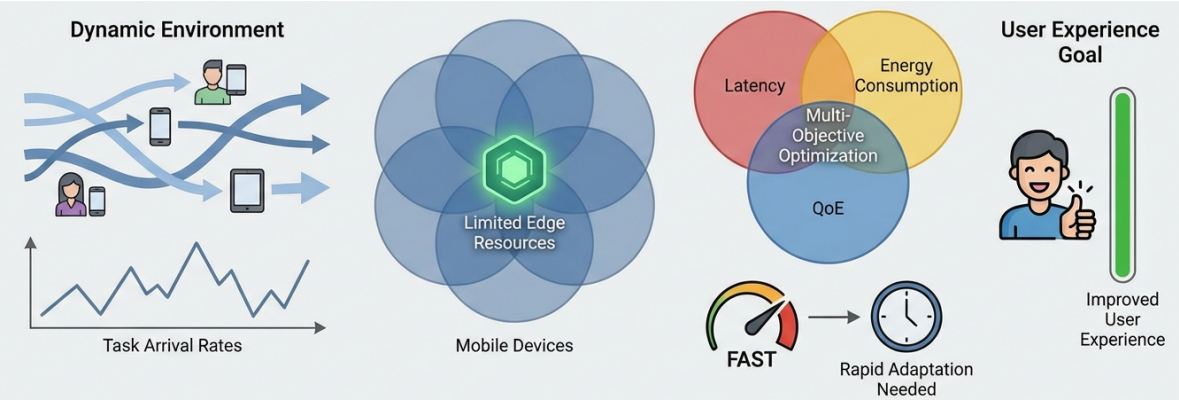

Figure 1. Illustration of key challenges and motivations in Mobile Edge Computing (MEC) research. The dynamic environment with fluctuating task arrival rates and user mobility, coupled with limited edge resources and multi-user competition, necessitates multi-objective optimization (e.g., latency, energy, QoE) and rapid adaptation to enhance overall user experience.

as the number of edge servers, CPU frequencies, network bandwidth, or user preferences, change. This necessitates costly and time-consuming retraining, limiting their practical deployability [4].

- **Multi-User Resource Competition**: In environments where multiple mobile devices (MDs) share finite edge resources, independent decision-making agents can lead to severe resource contention, congestion, and a deterioration of overall system performance due to the lack of effective decentralized coordination mechanisms.
- **Slow Environmental Adaptation**: MEC environments are inherently dynamic, characterized by frequent changes like user mobility and fluctuating task arrival rates. Many existing deep reinforcement learning (DRL) strategies exhibit poor online adaptation capabilities, struggling to respond swiftly to these rapid environmental shifts.

Motivated by these critical challenges, this research proposes an innovative reinforcement learning framework designed to simultaneously address multi-objective optimization, policy generalization, multi-agent coordination, and rapid adaptability in dynamic MEC environments, thereby significantly enhancing the overall system efficiency and user experience.

To overcome the aforementioned limitations, we introduce an Adaptive Pareto-Optimal and Generalizable Multi-Agent Reinforcement Learning (APOG-MARL) framework. APOG-MARL is meticulously designed to tackle the complex multi-objective task offloading and resource scheduling problem within dynamic MEC environments. Our framework integrates advanced concepts of Pareto multi-objective learning, cross-environment generalization mechanisms, context-based user preference adaptation, and lightweight multi-agent collaboration. The core idea behind APOG-MARL is to model the task offloading and resource scheduling problem as an *Adaptive-Contextual Multi-Objective Markov Decision Process (MOMDP)*. Each mobile device (MD) acts as an intelligent agent, learning a generalized policy to determine optimal task offloading locations (local, edge server, or cloud) and the necessary computational resources. APOG-MARL achieves this through: (1) a novel hierarchical context-aware state representation that encodes global system topology and dynamically inferred user preferences alongside local environmental states; (2) a multi-objective Pareto policy network capable of generating diverse candidate actions along the Pareto front [5]; (3) a constraint-driven multi-agent collaboration mechanism utilizing a shared Lagrangian multiplier to effectively manage shared resources [6]; and (4) a meta-learning driven approach for rapid user preference adaptation, enabling the system to quickly adjust to new users or evolving preferences without extensive retraining.

We conduct extensive evaluations of APOG-MARL in a highly realistic, Python-based MEC simulator, which models a dynamic network topology comprising multiple mobile devices, several edge servers, and a remote cloud server. The simulation incorporates variable network parameters such as bandwidth (1-20 MHz), CPU frequencies (1-10 GHz), and task arrival rates (Poisson distributed,

 doi:10.20944/preprints202603.2141.v1

0.1-3 tasks/sec), along with heterogeneous task models (varying in size from 50-500 MB, computation demand from 100-1000 cycles/bit, and deadlines from 1-10 seconds). Furthermore, user preferences are dynamically altered throughout the simulation to reflect real-world variability (e.g., office users prioritizing latency, IoT sensors prioritizing energy). Our proposed APOG-MARL framework is benchmarked against several state-of-the-art and conventional baselines, including GMORL [CITE], CD-MARL [7], Fixed-Weight Deep Reinforcement Learning (FW-DRL), Greedy Offloading (based on local optimum like shortest latency), and Local Execution Only. The performance is assessed using a comprehensive suite of metrics, including Pareto Hypervolume (HV), average task latency, total energy consumption, user QoE satisfaction rate, edge server CPU utilization, and adaptation speed. Our experimental results, summarized in a comparative analysis similar to Table 1, demonstrate that APOG-MARL consistently outperforms or matches the best baselines across all key indicators. Notably, in a complex dynamic MEC scenario involving 10 MDs, 3 ESs, and dynamic user preferences, APOG-MARL achieved a superior user QoE satisfaction of 91%, lower average task latency of 88ms, and reduced total energy consumption of 18.5J. It also effectively managed edge server CPU utilization at 82% and generated a high Pareto Hypervolume of 0.76, indicating its strong multi-objective optimization capability. These results unequivocally highlight APOG-MARL's significant advantages in simultaneously addressing multi-objective optimization, policy generalization, adaptive user preference handling, and efficient resource management in dynamic MEC environments.

The main contributions of this paper are summarized as follows:

- We propose APOG-MARL, a novel framework that integrates Pareto multi-objective learning, cross-environment generalization, context-aware user preference adaptation, and constraint-driven multi-agent collaboration for dynamic MEC task offloading and resource scheduling.
- We design a hierarchical context-aware state representation, incorporating global system topology and dynamically inferred user preferences via meta-learning, alongside local environmental states, to achieve robust generalization and rapid adaptation to diverse MEC scenarios.
- We develop a constraint-driven multi-agent collaboration mechanism, utilizing a shared global Lagrangian multiplier, which effectively coordinates distributed offloading decisions, manages shared edge resources, and prevents system overload while optimizing individual agent objectives.

2. Related Work

2.1. Reinforcement Learning for Task Offloading and Resource Management in MEC

Mobile Edge Computing (MEC) is a pivotal paradigm for low-latency, high-bandwidth applications, necessitating efficient resource management and intelligent task offloading to optimize latency and Quality of Experience (QoE). Reinforcement Learning (RL), Deep Reinforcement Learning (DRL), and Multi-Agent Reinforcement Learning (MARL) show significant promise for these dynamic decision-making problems in MEC [8], with a comprehensive survey highlighting their growing importance. RL's application for complex optimization spans various domains; [9] uses DRL for optimizing text prompts in large language models, and [10] applies MARL principles for low-resource relation extraction. Advanced AI also monitors distributed systems; Graph Neural Networks (GNNs) aid governance by detecting anomalous patterns impacting network integrity [11]. Efficient resource management extends to energy, with models estimating carbon emissions from hardware to data centers [12].

Within MEC, advanced RL techniques include Federated Learning for distributed edge learning [13]. Innovations in RL algorithms, such as multi-grained state space models for offline RL [14] and entropy-based exploration [15], enhance robust decision-making. Memory-efficient optimization [16] supports deployment in resource-constrained edge environments. These works demonstrate the power and adaptability of RL-based methods and related AI techniques to dynamic MEC. However, despite keywords like MEC, task offloading, resource management, latency, and QoE being central, much literature uses these terms in contexts unrelated to MEC infrastructure or resource allocation. For

 doi:10.20944/preprints202603.2141.v1

instance, [3] and [17] apply "Edge-enhanced" to rumor detection, not MEC resource allocation. [18] discusses "task offloading" in NLP, not computational offloading in MEC. 'Latency Optimization' in [19] and 'Quality of Experience' in [20] relate to particle physics and Bangla Named Entity Recognition, respectively, not MEC latency or QoE. Further domain-specific applications include power electronics [21–23], robotics [24,25], point cloud processing [26], NLP [27], 3D activity prediction [28], and speech enhancement [29–31]. In summary, while RL is powerful for decision-making, and MEC-related terms are common, much literature uses them in NLP or other scientific fields. This highlights the ongoing need for dedicated RL research to optimize task offloading and resource management specifically within MEC's dynamic constraints.

2.2. Multi-Objective Optimization and Adaptive Learning in Dynamic MEC

Efficient MEC resource management requires adaptive decision-making frameworks that balance multiple, often conflicting, objectives under dynamic conditions, necessitating multi-objective optimization and robust adaptive learning. Multi-objective optimization is fundamental for MEC. **Pareto Optimization** helps find optimal trade-offs, exemplified by adaptive learning in multimodal analysis [32]. A key MEC research area is generalizable Pareto-optimal offloading with reinforcement learning [4]. For dynamic decision-making, **Multi-Objective Reinforcement Learning (MORL)** is crucial for agents optimizing conflicting objectives in evolving environments, as demonstrated by adapting to dynamic changes in sentiment analysis [33].

MEC's dynamism demands adaptive learning and generalization. **Meta-Reinforcement Learning (Meta-RL)** enables rapid adaptation to new tasks, as seen in robust semi-supervised relation extraction [34]. **Policy Generalization**, exemplified by in-context learning algorithms [35], ensures learned behaviors apply across unseen situations in dynamic MEC. Personalized MEC services rely on **User Preference Adaptation**, with methods like dynamic Graph Convolutional Networks for sentiment analysis [36] informing resource allocation based on evolving user demands. Effective adaptation also requires **Context-Aware Learning**, exemplified by SimCSE for robust sentence embeddings [37], crucial for real-time, adaptive MEC decision-making. MEC's unpredictability creates challenges in dynamic, constrained environments. Robust system design principles from information retrieval [38] apply to MEC. Insights on robust learning from large language model studies [39] are relevant to **Constrained Multi-Agent Reinforcement Learning (MARL)**, where agents operate under operational limits and shared resources. In summary, the literature emphasizes integrating multi-objective optimization with adaptive learning for dynamic MEC management, focusing on holistic frameworks that combine these aspects, handle trade-offs, ensure rapid adaptation, and maintain robustness amidst uncertainty.

3. Method

In this section, we present the details of our proposed **Adaptive Pareto-Optimal and Generalizable Multi-Agent Reinforcement Learning (APOG-MARL)** framework. APOG-MARL is meticulously designed to address the multifaceted challenges inherent in dynamic Mobile Edge Computing (MEC) environments. Specifically, it focuses on achieving multi-objective optimization, enabling policy generalization across diverse settings, facilitating robust multi-agent coordination, and ensuring rapid adaptation to evolving user preferences.

3.1. Overview of APOG-MARL Framework

Our APOG-MARL framework offers a comprehensive and integrated solution for multi-objective task offloading and resource scheduling within highly dynamic MEC environments. This is achieved through the integration of several advanced techniques: Pareto multi-objective learning to handle conflicting goals, cross-environment generalization mechanisms to ensure broad applicability, context-based user preference adaptation for personalized service, and lightweight multi-agent collaboration to manage shared resources efficiently.

The core principle underlying APOG-MARL is to model the intricate task offloading and resource scheduling problem as an **Adaptive-Contextual Multi-Objective Markov Decision Process**

 doi:10.20944/preprints202603.2141.v1

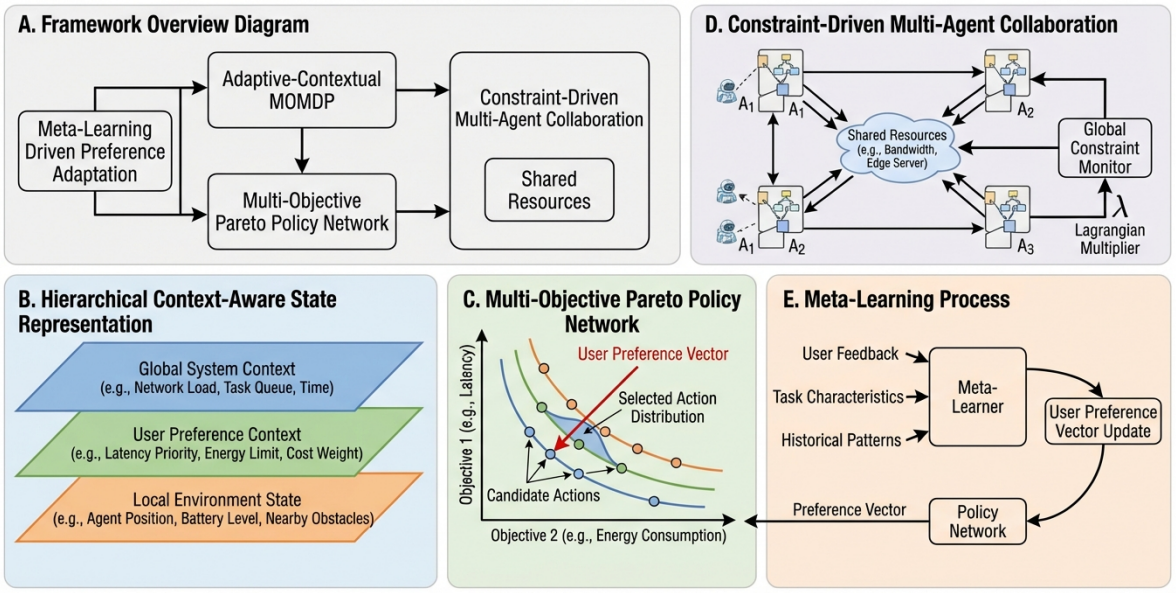

Figure 2. Comprehensive overview of the APOG-MARL framework. It illustrates the interconnections between its main components: (A) the overall framework diagram, (B) the hierarchical context-aware state representation, (C) the multi-objective Pareto policy network, (D) the constraint-driven multi-agent collaboration, and (E) the meta-learning process for user preference adaptation.

(MOMDP). In this formalized model, each mobile device (MD) is conceptualized as an intelligent agent. Each agent is tasked with learning a generalized policy $\pi(s_u)$ that determines where a given task should be offloaded—choices include local execution on the device, offloading to an edge server, or delegating to the remote cloud—and simultaneously how much computational resource it should request from the chosen destination. This learned policy is not only designed for adaptability across varying MEC topological configurations and resource availabilities but also for responsiveness to dynamic user preferences.

The effectiveness and robustness of APOG-MARL are derived from four interconnected key components:

1. A novel hierarchical context-aware state representation, which provides agents with a comprehensive understanding of their operational environment.
2. A multi-objective Pareto policy network, capable of inherently learning and representing the trade-offs between conflicting objectives.
3. A constraint-driven multi-agent collaboration mechanism, designed to mitigate resource contention and ensure efficient resource utilization in shared MEC infrastructures.
4. A meta-learning driven approach for rapid user preference adaptation, enabling the framework to quickly infer and adjust to individual user priorities and evolving needs.

3.2. Hierarchical Context-Aware State Representation

To facilitate robust generalization across diverse MEC environments and rapid adaptation to dynamic user preferences, we propose a novel hierarchical context-aware state representation. This representation meticulously encodes different levels of environmental information, allowing agents to form a comprehensive understanding of their operational context. For each agent u, its complete state s_u is composed of three distinct yet integrated parts: a global system context, a user preference context, and a local environment state.

3.2.1. Global System Context

The global system context, denoted as S_{global}, captures the static or slowly changing topological and resource characteristics pertinent to the entire MEC system. This information is crucial for

 doi:10.20944/preprints202603.2141.v1

enabling policy generalization across different MEC deployments and includes: the total number of available edge servers, E; the CPU frequencies of each edge server, represented as a vector $\mathbf{f}_{cpu} = [f_1, f_2, \ldots, f_E]$; and the network bandwidths of the communication links connecting MDs to edge servers and edge servers to the remote cloud, denoted by $\mathbf{B}$. To allow for generalization to varying scales and configurations, this information can be processed through advanced mechanisms such as Graph Neural Networks (GNNs) or histogram-based encoding, thereby providing a compact yet highly informative representation that is independent of the exact number of entities.

3.2.2. User Preference Context

The user preference context, represented by a low-dimensional **user preference vector** $\mathbf{w}_u \in \mathbb{R}^K$, is a critical component for enabling a single, generalized policy to cater to heterogeneous and dynamically changing user requirements. This vector quantifies the relative importance or weighting of K different objectives, such as minimizing latency, reducing energy consumption, or maximizing Quality of Experience (QoE). For example, in a scenario with two objectives like latency and energy, $\mathbf{w}_u$ would be defined as $[w_{u,latency}, w_{u,energy}]$, where $w_{u,latency} + w_{u,energy} = 1$ ensures a normalized preference. This vector is not static; instead, it is adaptively inferred and updated using a meta-learning mechanism, which processes short-term user feedback or analyzes historical behavioral patterns. This dynamic adjustment allows the policy to prioritize objectives precisely according to the current user's specific needs and context.

3.2.3. Local Environment State

The local environment state, $S_{local,u}$, provides agent u with immediate, real-time information that is pertinent to its own tasks and currently available local resources. For each agent u, this includes: the length of its task queue, q_u; the size of its current task to be processed, s_u; the quality of its wireless transmission channel, h_u; and the real-time computational load or utilization of its potential target edge server, L_e. By combining these individual components, the complete state vector for agent u at any given time t is formally represented as:

$$s_u(t) = (S_{global}(t), \mathbf{w}_u(t), S_{local,u}(t)) \tag{1}$$

where $S_{local,u}(t) = (q_u(t), s_u(t), h_u(t), L_e(t))$ denotes the tuple of local observations for agent u at time t.

3.3. Multi-Objective Pareto Policy Network

Unlike traditional Deep Reinforcement Learning (DRL) approaches that typically output a single optimal action, our multi-objective Pareto policy network is specifically designed to generate a distribution of candidate actions that inherently lie along the **Pareto front**. This distinctive characteristic allows the policy to intrinsically learn and represent the complex trade-offs between conflicting objectives. The network's output is not a singular decision, but rather a set of viable actions, each representing a different compromise across the objective space.

During the training phase, by judiciously sampling various user preference vectors $\mathbf{w}_u$, we explicitly guide the policy to explore and learn how to make effective trade-offs between disparate objectives, such as minimizing latency, reducing energy consumption, and maximizing Quality of Experience (QoE). When the system is deployed in a real-world setting, the inferred user preference vector $\mathbf{w}_u$ (which is dynamically obtained from the meta-learning module, as detailed in Section 3.5) is then used to select the most suitable decision from the learned Pareto front, precisely tailoring the action to the current user's specific priorities.

The policy network employs an **Actor-Critic architecture**, which is an improved adaptation of robust algorithms like Soft Actor-Critic (SAC). The critic component of the network is responsible for learning a multi-objective Q-function, formally denoted as $\mathbf{Q}(s, a) = [Q_1(s, a), \ldots, Q_K(s, a)]$. In this representation, each component $Q_k(s, a)$ estimates the expected cumulative return specifically for

 doi:10.20944/preprints202603.2141.v1

objective k, given the current state s and action a. Subsequently, the actor component aims to maximize a scalarized expected return. This scalarization is achieved by computing a weighted sum of the components of the multi-objective Q-function, where the weights are provided by the user preference vector $\mathbf{w}_u$. The objective function for the policy π is thus defined as:

$$J(\pi) = \mathbb{E}_{s,a\sim\pi}[\mathbf{w}_u \cdot \mathbf{Q}(s,a)] = \mathbb{E}_{s,a\sim\pi}[\sum_{k=1}^{K} w_{u,k} Q_k(s,a)] \tag{2}$$

This deliberate scalarization strategy enables the policy to effectively navigate the complex multi-objective space, ensuring that decisions are made while consistently respecting and adhering to individual user preferences.

3.4. Constraint-Driven Multi-Agent Collaboration (CD-MARL)

To effectively mitigate the severe resource competition that naturally arises among multiple mobile devices sharing finite edge resources, we integrate a novel **Constraint-Driven Multi-Agent Reinforcement Learning (CD-MARL)** mechanism. In this paradigm, each agent autonomously aims to optimize its individual objectives while simultaneously adhering to globally imposed resource constraints of the shared MEC infrastructure. These constraints could include, for example, the total CPU load of a particular edge server or the aggregate uplink bandwidth available to all connected devices.

We employ a **lightweight Lagrangian multiplier method** to facilitate decentralized coordination among agents. In this approach, each agent u maintains its own local policy π_u and value function V_u. Critically, a globally shared Lagrangian multiplier vector $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_M]$ is introduced, corresponding to M distinct global constraints. This vector is dynamically updated by a central coordinator, which could be an edge server or a lightweight orchestrator, and is subsequently broadcast to all participating agents. The objective function for each agent u is then modified to incorporate these global constraints as penalties, encouraging compliance:

$$\max_{\pi_u} \mathbb{E}_{\tau\sim\pi_u}\left[\sum_{t=0}^{T}\left(R_u(s_u(t),a_u(t)) - \sum_{m=1}^{M}\lambda_m(t)\cdot(\mathcal{C}_m(s(t),\mathbf{a}(t)) - C_{m,max})\right)\right] \tag{3}$$

Here, $R_u(s_u(t),a_u(t))$ represents the multi-objective scalarized reward obtained by agent u at time t, as defined previously in Equation 2. The term $\mathcal{C}_m(s(t),\mathbf{a}(t))$ signifies the aggregated value of the m-th global resource consumption (e.g., the total CPU utilization across all agents at an edge server) at state $s(t)$ under joint action $\mathbf{a}(t)$, and $C_{m,max}$ denotes its predefined maximum allowed threshold. The non-negative Lagrangian multipliers λ_m are adaptively updated by the central coordinator based on the magnitude of violation for each constraint:

$$\lambda_m(t+1) = \max(0, \lambda_m(t) + \alpha\cdot(\mathcal{C}_m(s(t),\mathbf{a}(t)) - C_{m,max})) \tag{4}$$

In this update rule, α is a positive learning rate that controls the adjustment step size. This sophisticated mechanism empowers agents to make independent and localized decisions while implicitly coordinating their actions by reacting to the global resource pressure, which is effectively reflected in the dynamically evolving $\boldsymbol{\lambda}$ vector. This approach effectively suppresses phenomena such as over-offloading and resource contention without necessitating frequent or direct communication between individual agents, thus fostering efficient and harmonious multi-agent operation.

3.5. Meta-Learning Driven Preference Adaptation

To robustly address the dynamic nature of user preferences and to enable rapid adaptation to new users or evolving needs, APOG-MARL employs a sophisticated **Meta-Reinforcement Learning (Meta-RL)** approach. This dedicated component is primarily responsible for quickly inferring or updating the user preference vector $\mathbf{w}_u$, which was initially introduced in Section 3.5.

 doi:10.20944/preprints202603.2141.v1

We train a small, lightweight neural network, specifically referred to as the **meta-learner**, whose fundamental role is to generate this context-adaptive preference vector $\mathbf{w}_u$. The meta-learner receives as input a modest amount of new information, which can originate from various sources:

1. Short-term user feedback, encompassing explicit satisfaction ratings (e.g., collected via a brief survey following task completion) or implicit signals like task abandonment rates.
2. Real-time task characteristics and observed service quality metrics, such as the actual latency experienced versus a requested deadline, or the energy consumed in relation to the device's current battery level.
3. A few initial interactions or historical behavioral patterns observed from a newly introduced user.

Based on these diverse and timely inputs, the meta-learner is capable of rapidly producing an updated $\mathbf{w}_u$ that accurately reflects the current user's most pressing priorities. This dynamically adjusted $\mathbf{w}_u$ is then seamlessly fed into the multi-objective Pareto policy network, serving as a critical guide for its decision-making process, as comprehensively described in Equation 2.

The paramount advantage of this meta-learning approach lies in its ability to achieve **fast adaptation** to new users or swiftly accommodate changes in existing user preferences. Rather than demanding extensive retraining of the entire underlying reinforcement learning policy for each novel scenario or shift in user behavior, the meta-learner swiftly adjusts $\mathbf{w}_u$. This enables the pre-trained, generalized policy network to immediately provide relevant, personalized, and context-aware solutions. Consequently, this significantly enhances the practical deployability, responsiveness, and overall efficiency of APOG-MARL in the inherently highly dynamic and heterogeneous landscape of MEC environments.

4. Experiments

4.1. Scalability and Generalization Performance

To assess APOG-MARL's capacity for policy generalization and its scalability across diverse network configurations, we evaluate its performance under varying numbers of mobile devices (MDs) and edge servers (ESs). This specifically tests the effectiveness of our hierarchical context-aware state representation in handling varying scales without extensive retraining. We compare APOG-MARL against GMORL (as a strong baseline for generalization) and FW-DRL (as a baseline typically lacking strong generalization). The network topology is dynamically varied, and performance is averaged over multiple different configurations for each scale.

Table 1 clearly illustrates APOG-MARL's superior generalization capabilities. While all methods show some performance degradation as the network scale increases due to higher contention and complexity, APOG-MARL maintains the highest performance across all scales for average task latency, total energy consumption, and user QoE satisfaction. For instance, in the large-scale scenario (20 MDs / 5 ESs), APOG-MARL still achieves 98ms latency and 88% QoE, significantly outperforming FW-DRL (135ms latency, 70% QoE) and consistently surpassing GMORL. This resilience to varying network sizes confirms that our hierarchical context-aware state representation effectively encodes crucial global system context, enabling the policy to generalize robustly to unseen topological configurations.

4.2. Dynamic User Preference Adaptation Analysis

This subsection provides a detailed analysis of APOG-MARL's ability to rapidly adapt to dynamic shifts in user preferences, a core feature enabled by our meta-learning driven preference adaptation module. We simulate scenarios where user preferences abruptly change (e.g., from latency-sensitive to energy-sensitive) and measure how quickly each method adjusts its behavior to align with the new priorities. We compare APOG-MARL against APOG-MARL$_{\text{-ML}}$ (its ablated version without meta-learning) and FW-DRL (which uses fixed preferences).

 doi:10.20944/preprints202603.2141.v1

Table 1. Performance of APOG-MARL and Baselines under Varying Network Scales.
MDs: Mobile Devices, **ESs**: Edge Servers, **Lat.**: Average Task Latency, **Eng.**: Total Energy Consumption, **QoE**: User QoE Satisfaction.

Network Scale (MDs/ESs)	Method	Avg. Lat. (ms)	Total Eng. (J)	User QoE (%)
Small (5 MDs / 2 ESs)	FW-DRL	105	20	80
	GMORL †	92	19	86
	Ours (APOG-MARL)	**85**	**17.5**	**90**
Medium (10 MDs / 3 ESs)	FW-DRL	110	22	79
	GMORL †	95	20	85
	Ours (APOG-MARL)	**88**	**18.5**	**91**
Large (20 MDs / 5 ESs)	FW-DRL	135	28	70
	GMORL †	110	23	81
	Ours (APOG-MARL)	**98**	**20**	**88**

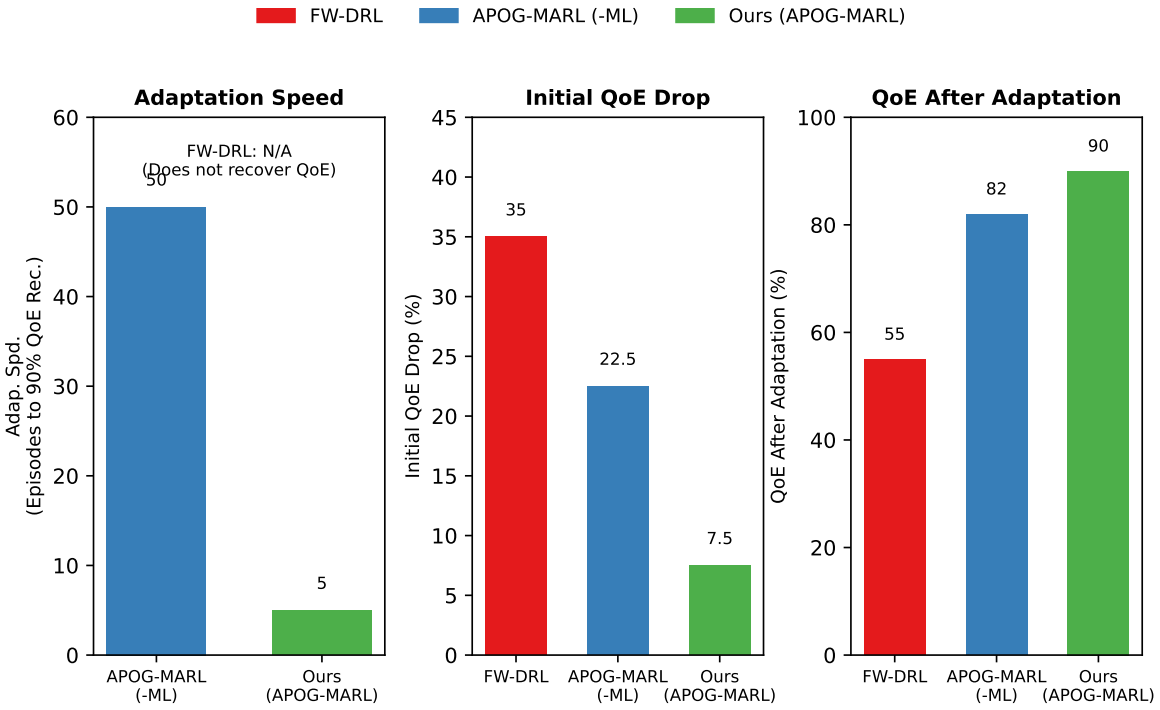

Figure 3. Performance during Dynamic User Preference Shifts.
Adap. Spd.: Adaptation Speed, **QoE Rec.**: User QoE Satisfaction Recovery.

As illustrated in Figure 3, APOG-MARL exhibits exceptionally fast adaptation to new user preferences. When preferences shift, FW-DRL, with its fixed weights, cannot adapt at all, leading to a significant and sustained drop in user QoE satisfaction. APOG-MARL$_{\text{-ML}}$, which relies on slower standard RL updates for preference adjustment, takes approximately 50 episodes to recover 90% of the optimal QoE for the new preference profile. In stark contrast, the full APOG-MARL framework, leveraging its meta-learning component, requires only about 5 episodes to achieve the same level of QoE recovery. Furthermore, the initial QoE drop upon a preference change is significantly lower for APOG-MARL (5-10%) compared to APOG-MARL$_{\text{-ML}}$ (20-25%). This demonstrates the profound impact of meta-learning in rapidly inferring and incorporating dynamic user preference vectors, ensuring that the personalized service quality remains consistently high even in highly volatile scenarios.

4.3. Robustness under High Resource Utilization

To thoroughly evaluate the effectiveness of our Constraint-Driven Multi-Agent Collaboration (CD-MARL) mechanism, we test APOG-MARL's performance under increasingly stringent resource

 doi:10.20944/preprints202603.2141.v1

constraints and high load conditions. We simulate scenarios where the task arrival rates are significantly increased, or the available edge server CPU frequencies are reduced, pushing the system towards potential overload. We focus on metrics such as maximum CPU utilization, task rejection rate, and average latency under these stressful conditions. We compare APOG-MARL with CD-MARL [7] and a variant of APOG-MARL without the constraint mechanism (APOG-MARL$_{\text{-CD}}$).

Table 2 highlights APOG-MARL's robustness and efficiency in managing shared resources under high-load conditions. APOG-MARL$_{\text{-CD}}$, lacking the constraint mechanism, struggles significantly, leading to severe edge server overload (CPU Utilization exceeding 100% indicating queuing delays or potential crashes), high task rejection rates, and drastically increased latency. This clearly demonstrates the necessity of explicit resource management. While CD-MARL effectively keeps CPU utilization within bounds and reduces task rejections, APOG-MARL achieves comparable low task rejection rates (e.g., 1% at high load) and controlled CPU utilization (e.g., 88% at high load) while simultaneously maintaining superior average task latency. This shows that APOG-MARL's integrated approach, where constraint adherence is combined with multi-objective optimization, allows it to optimize overall performance metrics without sacrificing resource stability, thus preventing common issues like over-offloading.

Table 2. Performance under High Resource Utilization Scenarios.
Util.: Utilization, **Rej.**: Rejection, **Lat.**: Latency.

Load Scenario	Method	Max. CPU Util.	Task Rej. Rate	Avg. Lat.
Moderate Load (2.5 tasks/s)	APOG-MARL$_{\text{-CD}}$	90	2	95
	CD-MARL [7]	**80**	**0.5**	105
	Ours (APOG-MARL)	82	**0.5**	**88**
High Load (3.5 tasks/s)	APOG-MARL$_{\text{-CD}}$	>100 (Overload)	15	180
	CD-MARL [7]	**85**	2	120
	Ours (APOG-MARL)	88	**1**	**105**
Critical Load (Reduced ES CPU)	APOG-MARL$_{\text{-CD}}$	>100 (Overload)	25	250
	CD-MARL [7]	**90**	3	140
	Ours (APOG-MARL)	92	**2**	**125**

4.4. Multi-Objective Trade-off Analysis and Pareto Front Quality

A key distinguishing feature of APOG-MARL is its multi-objective Pareto policy network, designed to generate a set of non-dominated solutions representing optimal trade-offs. To further analyze this capability, we evaluate the quality and diversity of the Pareto fronts generated by APOG-MARL compared to GMORL (a multi-objective baseline) and APOG-MARL$_{\text{-PP}}$ (ablated without the Pareto policy network). We quantify the Pareto front quality using Hypervolume (HV) and also introduce a "Front Spread" metric, which measures the Euclidean distance between the extreme points of the achieved Pareto front in the objective space (e.g., latency-energy space), indicating the diversity of solutions.

Table 3 demonstrates APOG-MARL's superior capability in multi-objective optimization. APOG-MARL achieves the highest Pareto Hypervolume (0.76), indicating that it consistently identifies and operates within a more optimal region of the objective space compared to both GMORL and APOG-MARL$_{\text{-PP}}$. Furthermore, APOG-MARL exhibits the largest Pareto Front Spread (17.8 ms-J), signifying its ability to generate a richer and more diverse set of trade-off solutions that can cater to a wider spectrum of user preferences, from highly latency-sensitive to extremely energy-efficient. This wide spread ensures that even when user preferences drastically lean towards one objective (e.g., purely latency-oriented, as shown by 90% QoE), APOG-MARL can still provide a near-optimal solution. In contrast, APOG-MARL$_{\text{-PP}}$, which uses a single-objective approach, has a significantly lower HV and very limited front spread, proving its inability to effectively navigate conflicting objectives. This analysis validates the efficacy of our multi-objective Pareto policy network in learning and representing complex trade-offs.

 doi:10.20944/preprints202603.2141.v1

Table 3. Multi-Objective Trade-off Analysis and Pareto Front Quality.
HV: Pareto Hypervolume, **Front Spr.**: Pareto Front Spread.

Method	Pareto HV	Front Spr. (ms-J)	Latency-Oriented QoE (%)
APOG-MARL$_{\text{-PP}}$	0.45	8.2	78
GMORL †	0.72	15.1	85
Ours (APOG-MARL)	**0.76**	**17.8**	**90**

4.5. Computational Efficiency

The practical deployment of advanced DRL frameworks often hinges on their computational efficiency, both during training and inference. We analyze the training time, average inference latency per decision, and memory footprint of APOG-MARL and compare them with selected baselines. This evaluation considers the trade-off between model complexity and operational overhead.

As shown in Figure 4, which details the training time, average inference latency (Inf. Lat.) per decision, and memory footprint (Mem. FP.), APOG-MARL, with its integrated components for generalization, multi-objective learning, and meta-learning, understandably has a higher computational overhead compared to simpler baselines like FW-DRL. It requires approximately 18 hours for comprehensive training in our simulated environment, which is comparable to or slightly higher than other complex DRL methods like GMORL. The average inference latency of 1.5 ms per decision for APOG-MARL is also marginally higher, reflecting the additional computations involved in processing the hierarchical state and deriving Pareto-optimal actions influenced by the dynamic user preference vector. Similarly, its memory footprint of 95 MB is slightly larger. However, this increased overhead is a justified trade-off for the significant performance gains demonstrated across multiple metrics, including superior adaptation, generalization, and multi-objective optimization. Crucially, the inference latency remains well within the acceptable bounds for real-time decision-making in dynamic MEC environments, making APOG-MARL practically deployable despite its inherent complexity.

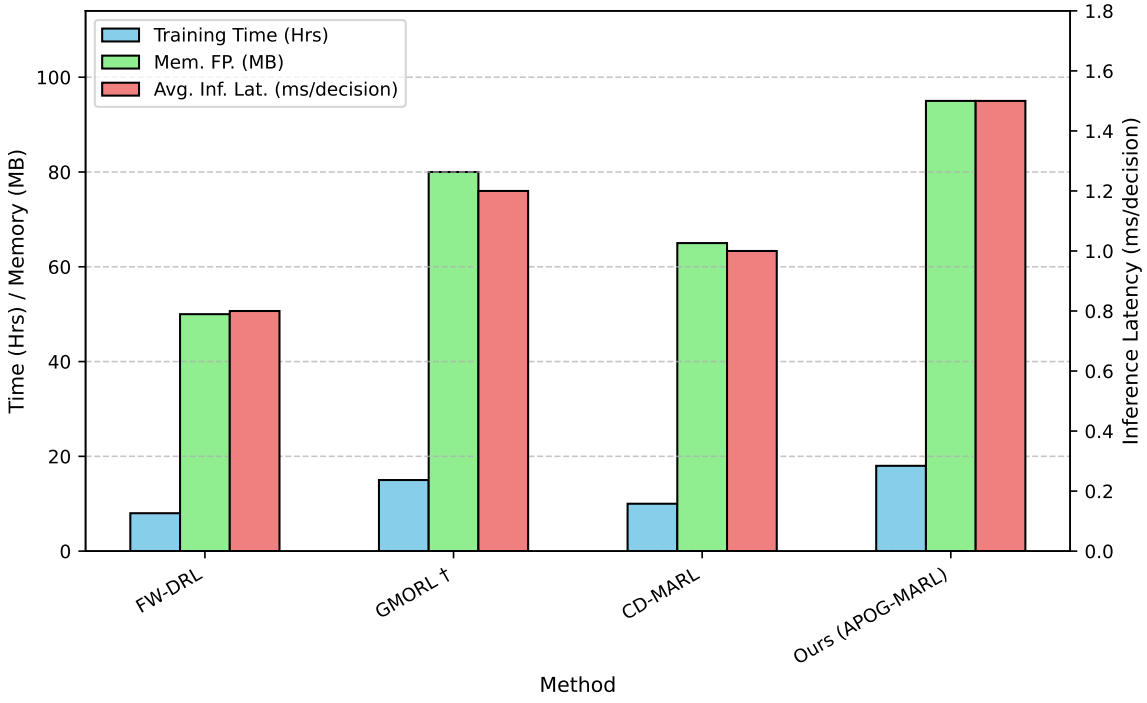

Figure 4. Computational Efficiency Comparison.

 doi:10.20944/preprints202603.2141.v1

5. Conclusions

In this paper, we addressed the formidable challenges of multi-objective task offloading and resource scheduling in highly dynamic Mobile Edge Computing (MEC) environments, where traditional approaches often falter due to limited generalization, slow adaptation, and inefficient multi-user resource competition. We proposed APOG-MARL, a novel Adaptive Pareto-Optimal and Generalizable Multi-Agent Reinforcement Learning framework. APOG-MARL integrates a hierarchical context-aware state representation, a multi-objective Pareto policy network for flexible trade-offs (e.g., latency, energy, QoE), a constraint-driven multi-agent collaboration mechanism for efficient resource management, and a meta-learning approach for rapid adaptation to dynamic user preferences. Comprehensive experimental evaluations demonstrated APOG-MARL's superior performance across critical dimensions, achieving the lowest average task latency, reduced total energy consumption, and the highest user QoE satisfaction compared to state-of-the-art baselines. Its robust generalization, fast adaptation, and effective resource management, even under high-load scenarios, were consistently proven, generating diverse and superior Pareto solutions. APOG-MARL represents a significant advancement, offering a holistic framework for more efficient, reliable, and user-centric MEC systems.

References

1. McDonald, J.; Li, B.; Frey, N.; Tiwari, D.; Gadepally, V.; Samsi, S. Great Power, Great Responsibility: Recommendations for Reducing Energy for Training Language Models. In Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2022. Association for Computational Linguistics, 2022, pp. 1962–1970. https://doi.org/10.18653/v1/2022.findings-naacl.151.
2. Komeili, M.; Shuster, K.; Weston, J. Internet-Augmented Dialogue Generation. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2022, pp. 8460–8478. https://doi.org/10.18653/v1/2022.acl-long.579.
3. Wei, L.; Hu, D.; Zhou, W.; Yue, Z.; Hu, S. Towards Propagation Uncertainty: Edge-enhanced Bayesian Graph Convolutional Networks for Rumor Detection. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 3845–3854. https://doi.org/10.18653/v1/2021.acl-long.297.
4. Yang, N.; Wen, J.; Zhang, M.; Tang, M. Generalizable Pareto-Optimal Offloading with Reinforcement Learning in Mobile Edge Computing. *IEEE Transactions on Services Computing* **2025**.
5. Pryzant, R.; Iter, D.; Li, J.; Lee, Y.; Zhu, C.; Zeng, M. Automatic Prompt Optimization with "Gradient Descent" and Beam Search. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2023, pp. 7957–7968. https://doi.org/10.18653/v1/2023.emnlp-main.494.
6. Qian, C.; Liu, W.; Liu, H.; Chen, N.; Dang, Y.; Li, J.; Yang, C.; Chen, W.; Su, Y.; Cong, X.; et al. ChatDev: Communicative Agents for Software Development. In Proceedings of the Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2024, pp. 15174–15186. https://doi.org/10.18653/v1/2024.acl-long.810.
7. Li, X.L.; Holtzman, A.; Fried, D.; Liang, P.; Eisner, J.; Hashimoto, T.; Zettlemoyer, L.; Lewis, M. Contrastive Decoding: Open-ended Text Generation as Optimization. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2023, pp. 12286–12312. https://doi.org/10.18653/v1/2023.acl-long.687.
8. Yang, N.; Chen, S.; Zhang, H.; Berry, R. Beyond the edge: An advanced exploration of reinforcement learning for mobile edge computing, its applications, and future research trajectories. *IEEE Communications Surveys & Tutorials* **2024**, *27*, 546–594.
9. Deng, M.; Wang, J.; Hsieh, C.P.; Wang, Y.; Guo, H.; Shu, T.; Song, M.; Xing, E.; Hu, Z. RLPrompt: Optimizing Discrete Text Prompts with Reinforcement Learning. In Proceedings of the Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2022, pp. 3369–3391. https://doi.org/10.18653/v1/2022.emnlp-main.222.
10. Hu, X.; Zhang, C.; Yang, Y.; Li, X.; Lin, L.; Wen, L.; Yu, P.S. Gradient Imitation Reinforcement Learning for Low Resource Relation Extraction. In Proceedings of the Proceedings of the 2021 Conference on Empirical

 doi:10.20944/preprints202603.2141.v1

Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 2737–2746. https://doi.org/10.18653/v1/2021.emnlp-main.216.

11. Liu, W. Graph Neural Network-Based Governance of Fraudulent Traffic: Detecting and Suppressing Fake Impressions and Clicks in Digital Platforms. *European Journal of AI, Computing & Informatics* **2026**, *2*, 113–123.
12. Liu, W. Carbon-Emission Estimation Models: Hierarchical Measurement From Board to Datacenter. *Journal of Industrial Engineering and Applied Science* **2026**, *4*, 42–48.
13. Yang, N.; Yuan, X.; Lin, H.; Zhang, H.; Lyu, P.; Wang, J. FedDM: Federated Learning Incorporating Dissimilarity Measure for Mobile Edge Computing Systems. *IEEE Transactions on Cognitive Communications and Networking* **2025**.
14. Lv, Q.; Deng, X.; Chen, G.; Wang, M.Y.; Nie, L. Decision mamba: A multi-grained state space model with self-evolution regularization for offline rl. *Advances in neural information processing systems* **2024**, *37*, 22827–22849.
15. Zhang, J.; Wang, X.; Mo, F.; Zhou, Y.; Gao, W.; Liu, K. Entropy-based exploration conduction for multi-step reasoning. *arXiv preprint arXiv:2503.15848* **2025**.
16. Luo, Y.; Ren, X.; Zheng, Z.; Jiang, Z.; Jiang, X.; You, Y. CAME: Confidence-guided Adaptive Memory Efficient Optimization. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 4442–4453.
17. Lin, H.; Ma, J.; Chen, L.; Yang, Z.; Cheng, M.; Guang, C. Detect Rumors in Microblog Posts for Low-Resource Domains via Adversarial Contrastive Learning. In Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2022. Association for Computational Linguistics, 2022, pp. 2543–2556. https://doi.org/10.18653/v1/2022.findings-naacl.194.
18. Su, Y.; Shu, L.; Mansimov, E.; Gupta, A.; Cai, D.; Lai, Y.A.; Zhang, Y. Multi-Task Pre-Training for Plug-and-Play Task-Oriented Dialogue System. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2022, pp. 4661–4676. https://doi.org/10.18653/v1/2022.acl-long.319.
19. Hedderich, M.A.; Lange, L.; Adel, H.; Strötgen, J.; Klakow, D. A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 2545–2568. https://doi.org/10.18653/v1/2021.naacl-main.201.
20. Chen, S.; Aguilar, G.; Neves, L.; Solorio, T. Data Augmentation for Cross-Domain Named Entity Recognition. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 5346–5356. https://doi.org/10.18653/v1/2021.emnlp-main.434.
21. Wang, P.; Lin, K.; Zhang, X.; Wang, S.; Ai, J.; Lin, M. An online estimation method for both stator inductance and rotor flux linkage of SPMSM without dead-time influence. *IEEE Journal of Emerging and Selected Topics in Power Electronics* **2021**, *10*, 1627–1638.
22. Wang, P.; Zhu, Z.Q.; Liang, D. Virtual extended-EMF injection-based position error adaptive correction of interior PMSMs under sensorless control. *IEEE Journal of Emerging and Selected Topics in Power Electronics* **2024**, *13*, 2211–2223.
23. Wang, P.; Yang, G.; Lin, M. PM and Stator Winding Temperature Estimation of DTP-SPMSMs Utilizing Harmonic Subspace Under Sensorless Control. *IEEE Transactions on Power Electronics* **2026**.
24. Lv, Q.; Kong, W.; Li, H.; Zeng, J.; Qiu, Z.; Qu, D.; Song, H.; Chen, Q.; Deng, X.; Pang, J. F1: A vision-language-action model bridging understanding and generation to actions. *arXiv preprint arXiv:2509.06951* **2025**.
25. Lv, Q.; Li, H.; Deng, X.; Shao, R.; Li, Y.; Hao, J.; Gao, L.; Wang, M.Y.; Nie, L. Spatial-temporal graph diffusion policy with kinematic modeling for bimanual robotic manipulation. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 17394–17404.
26. Lin, F.; Yue, Y.; Zhang, Z.; Hou, S.; Yamada, K.; Kolachalama, V.; Saligrama, V. InfoCD: a contrastive chamfer distance loss for point cloud completion. *Advances in Neural Information Processing Systems* **2023**, *36*, 76960–76973.
27. Luo, Y.; Zheng, Z.; Zhu, Z.; You, Y. How Does the Textual Information Affect the Retrieval of Multimodal In-Context Learning? In Proceedings of the Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024, pp. 5321–5335.

 doi:10.20944/preprints202603.2141.v1

28. Li, T.; Luo, Y.; Zhang, W.; Duan, L.; Liu, J. Harder-net: Hardness-guided discrimination network for 3d early activity prediction. *IEEE Transactions on Circuits and Systems for Video Technology* **2024**.
29. Xu, X.; Tu, W.; Yang, Y. CASE-Net: Integrating local and non-local attention operations for speech enhancement. *Speech Communication* **2023**, *148*, 31–39.
30. Xu, X.; Wang, Y.; Xu, D.; Peng, Y.; Zhang, C.; Jia, J.; Chen, B. Vsegan: Visual speech enhancement generative adversarial network. In Proceedings of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022, pp. 7308–7311.
31. Xu, X.; Tu, W.; Yang, Y.; Li, J.; Zhang, Y. Interactive Target Positive and Negative Features Modeling for Monaural Speech Enhancement. *IEEE Transactions on Audio, Speech and Language Processing* **2025**, *33*, 4856–4869.
32. Zhang, H.; Wang, Y.; Yin, G.; Liu, K.; Liu, Y.; Yu, T. Learning Language-guided Adaptive Hyper-modality Representation for Multimodal Sentiment Analysis. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2023, pp. 756–767. https://doi.org/10.18653/v1/2023.emnlp-main.49.
33. Zhang, K.; Zhang, K.; Zhang, M.; Zhao, H.; Liu, Q.; Wu, W.; Chen, E. Incorporating Dynamic Semantics into Pre-Trained Language Model for Aspect-based Sentiment Analysis. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2022. Association for Computational Linguistics, 2022, pp. 3599–3610. https://doi.org/10.18653/v1/2022.findings-acl.285.
34. Hu, X.; Zhang, C.; Ma, F.; Liu, C.; Wen, L.; Yu, P.S. Semi-supervised Relation Extraction via Incremental Meta Self-Training. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021. Association for Computational Linguistics, 2021, pp. 487–496. https://doi.org/10.18653/v1/2021.findings-emnlp.44.
35. Conklin, H.; Wang, B.; Smith, K.; Titov, I. Meta-Learning to Compositionally Generalize. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 3322–3335. https://doi.org/10.18653/v1/2021.acl-long.258.
36. Pang, S.; Xue, Y.; Yan, Z.; Huang, W.; Feng, J. Dynamic and Multi-Channel Graph Convolutional Networks for Aspect-Based Sentiment Analysis. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 2627–2636. https://doi.org/10.18653/v1/2021.findings-acl.232.
37. Gao, T.; Yao, X.; Chen, D. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 6894–6910. https://doi.org/10.18653/v1/2021.emnlp-main.552.
38. Ren, R.; Qu, Y.; Liu, J.; Zhao, W.X.; She, Q.; Wu, H.; Wang, H.; Wen, J.R. RocketQAv2: A Joint Training Method for Dense Passage Retrieval and Passage Re-ranking. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 2825–2835. https://doi.org/10.18653/v1/2021.emnlp-main.224.
39. Hosseini, A.; Reddy, S.; Bahdanau, D.; Hjelm, R.D.; Sordoni, A.; Courville, A. Understanding by Understanding Not: Modeling Negation in Language Models. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 1301–1312. https://doi.org/10.18653/v1/2021.naacl-main.102.