Artificial Intelligence and Computing Innovations

Intelligence Science Publishing

https://www.ispublishing.com

ISSN: 3070-4944

Vo.4, No.4, 2024

Adaptive Scheduling for Multi-Model Collaborative Distributed Inference under Resource Heterogeneity and Dynamic Workloads

Sijia Li

University of Michigan, Ann Arbor, USA

sijialee@umich.edu

Abstract: This paper addresses the scheduling complexity of distributed inference systems in multi-model collaborative scenarios under resource heterogeneity, model diversity, and dynamic request conditions. An adaptive scheduling method for distributed inference systems with multi-model collaboration is proposed. From a system-level perspective, the method integrates model selection and resource allocation into a unified scheduling decision process. Coordinated management of multi-model inference workflows is achieved through comprehensive characterization of system states. The proposed scheduling mechanism dynamically adjusts its decisions in response to changes in the operating environment and maintains a reasonable execution order under shared resource constraints, thereby improving overall system performance. Comparative experiments conducted in multi-model collaborative inference settings demonstrate that the proposed method achieves clear advantages in resource utilization efficiency, load distribution rationality, system stability, and scheduling responsiveness. The method effectively mitigates performance degradation caused by resource contention and model heterogeneity during parallel multi-model execution. The results indicate that a unified adaptive scheduling design plays an important role in supporting complex intelligent inference services and provides valuable guidance for the engineering implementation and optimization of distributed inference systems.

Keywords: distributed inference systems; multi-model collaboration; adaptive scheduling; system performance optimization

1. Introduction

As artificial intelligence applications continue to evolve, the deployment paradigm in which a single model supports a single task has become insufficient for complex service scenarios. In practical systems,

functions such as search, recommendation, perception, and analysis often operate simultaneously. These functions are supported by models with different architectures, scales, and computational characteristics[1]. As a result, distributed inference systems increasingly exhibit the coexistence and collaborative execution of multiple models. In this context, inference requests are no longer mapped to a fixed model. Instead, they require flexible scheduling and coordinated execution across multiple models. This shift enhances system expressiveness and service coverage. At the same time, it places higher demands on scheduling mechanisms, making traditional single model-oriented designs difficult to apply directly.

While multi-model collaborative inference improves functional diversity, it also introduces more complex runtime behaviors[2]. On one hand, different models vary significantly in computational cost, memory usage, parallelism, and hardware affinity. Their inference processes, therefore, impose highly heterogeneous resource demands. On the other hand, models are often coupled through execution order, data dependencies, or shared intermediate results. Inference workflows thus evolve from independent request processing into complex processes that involve both cooperation and competition. Under high concurrency and dynamic workloads, the absence of effective scheduling amplifies resource contention and interference among models. This leads to increased latency variability, reduced overall throughput, and potential degradation of system stability and predictability.

Most existing distributed inference systems are designed around single model assumptions[3]. They typically treat inference tasks as independent and rely on static rules or heuristic parameters for request allocation. Such approaches can be effective when the number of models is small and workload patterns are stable. However, their limitations become evident in multi-model collaborative environments. Static scheduling cannot promptly reflect changes in model execution states or system resource conditions. It fails to address load imbalance and performance diversity across models. Moreover, viewing multi-model inference as a simple aggregation of multiple single model systems overlooks potential collaboration and global optimization opportunities. As a result, overall system performance is often underutilized.

With the continuous expansion of intelligent services, distributed inference systems now operate in increasingly dynamic and uncertain environments. Request arrival patterns exhibit strong temporal correlations and burstiness[4]. Model invocation frequencies change rapidly with business demands. System resource states are also affected by multi-tenant sharing, hardware heterogeneity, and environmental disturbances. In multi-model collaborative settings, such dynamics are further amplified. Scheduling decisions must jointly consider model selection, execution order, and resource allocation. Optimization from a single dimension or local perspective is insufficient to achieve global efficiency. There is a strong need for adaptive methods that can perceive system state changes, understand collaborative model behavior, and dynamically adjust scheduling strategies to support efficient inference services.

From a research perspective, adaptive scheduling for multi-model collaborative distributed inference systems represents a critical response to the increasing complexity of intelligent applications. It also offers new insights into the evolution of distributed system design. Systematic investigation of scheduling in multi-model inference can reveal intrinsic relationships between model heterogeneity and system resources. It can promote a transition from static configuration to dynamic and adaptive control. Progress in this direction provides theoretical foundations and methodological support for building efficient, flexible, and scalable inference infrastructures. This is essential for the stable operation and sustainable development of diverse intelligent services.

2. Related Work

Research on distributed inference systems has long focused on the efficient execution of deep learning inference in large-scale computing environments. Early studies mainly centered on single model services and addressed fundamental issues such as request scheduling, load balancing, and resource isolation. These works commonly adopted rule-based or heuristic approaches. Requests were allocated using predefined thresholds, priorities, or simple performance models to reduce latency and improve resource utilization. Although such

methods achieved reasonable performance in scenarios with limited model scale and low service complexity, they often relied on static assumptions about system behavior. As a result, they struggled to adapt to frequently changing workloads and resource conditions in real environments[5]. This limitation became more pronounced as the number of deployed models increased, leading to reduced scalability and robustness of scheduling strategies.

With the continuous evolution of artificial intelligence services, research attention gradually shifted toward the system challenges introduced by parallel deployment and the coexistence of multiple models. Some studies approached this problem from an architectural perspective. They explored techniques such as model partitioning, pipeline execution, and parallel inference to improve overall throughput. These methods alleviated single model bottlenecks to some extent. However, they typically treated model execution as a fixed computational unit, and scheduling decisions remained largely statically configured around model instances. Other works introduced finer-grained resource management mechanisms in multi-model settings. They aimed to reduce interference by limiting resource quotas or isolating different services[6]. Such approaches primarily focused on resource allocation itself. They provided limited insight into potential collaboration and dynamic interactions among models. As a result, achieving globally optimal scheduling under complex workloads remained difficult.

In recent years, learning based methods have been increasingly applied to complex decision-making problems. This trend has motivated studies that incorporate learning strategies into the scheduling and management of distributed systems. These approaches adjust scheduling decisions through online or offline learning based on system feedback. They are designed to handle uncertainty and non-stationarity in the environment. In distributed inference systems, learning driven methods have been used to predict request characteristics, estimate execution costs, or select suitable computing nodes. This improves system adaptivity. However, most existing studies focus on single model or homogeneous model scenarios. The scheduling objectives are relatively simple. The complex decision space arising from multi-model collaboration is often insufficiently considered. When multiple models with diverse computational characteristics and resource demands coexist, naive extensions of single model learning strategies fail to capture inter-model interactions effectively.

Overall, existing research has accumulated substantial theoretical and practical insights into scheduling for distributed inference systems. Yet clear gaps remain in adaptive scheduling for multi-model collaboration. On the one hand, many studies lack systematic modeling of collaborative model behavior. This limits the ability to exploit complementarity and sharing potential among models during scheduling. On the other hand, scheduling and resource management are often treated separately. Their strong coupling in multi-model environments is largely overlooked. As intelligent services continue to grow in scale and complexity, coordinating multi-model inference within a unified framework and enabling adaptive scheduling under dynamic conditions remain open challenges[7]. These issues represent critical problems that demand further investigation in the field of distributed inference systems.

3. Proposed Framwork

3.1 Overall Framework Description

Distributed inference systems designed for multi-model collaboration typically consist of multiple models with varying functionalities and computational characteristics, supporting complex intelligent service requirements. The overall model architecture is shown in Figure 1.

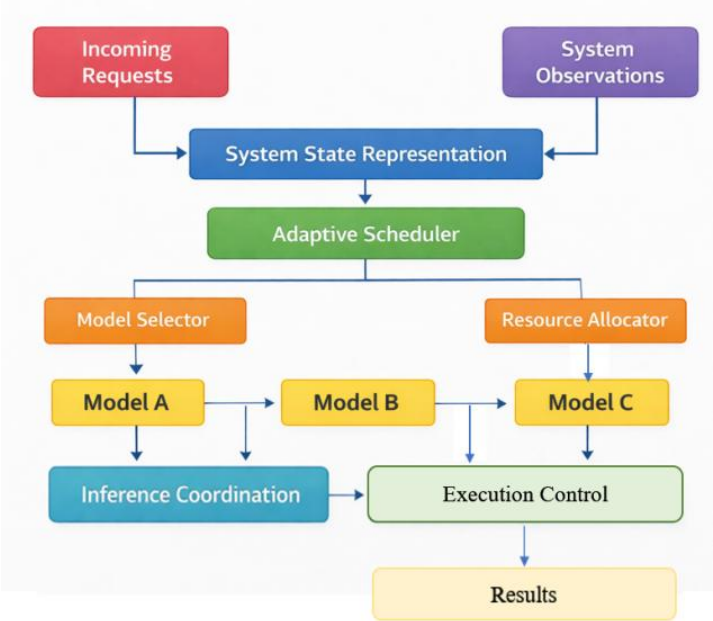

Figure 1. Overview of the proposed method

During operation, the system receives inference request streams from external sources and dynamically determines model selection, execution order, and resource allocation strategies based on request characteristics and system state. To characterize this process, the overall state of the system at time step t can be represented as a joint state vector, describing the model's running state, resource consumption, and request queue characteristics, thus providing a unified input for subsequent scheduling decisions:

$$s_t = [m_t, r_t, q_t] \tag{1}$$

Where m_t represents the running status of each model, r_t represents the system's available computing resources status, and q_t represents the current request queue characteristics. Based on the system status, the scheduling module outputs a scheduling action at each time step, describing the allocation of requests among multiple models and the resource orchestration method:

$$a_t = \pi(s_t) \tag{2}$$

Here, $\pi(\cdot)$ represents the scheduling policy function, whose goal is to achieve efficient collaborative execution of the multi-model inference process while satisfying system constraints. This framework lays the foundation for subsequent adaptive policy design by formalizing the multi-model scheduling problem as a mapping from state to action.

3.2 Multi-model collaborative reasoning, modeling, and resource constraint representation

In multi-model collaborative inference scenarios, different models have significantly different computational resource requirements. To uniformly describe model execution behavior, a model-level resource consumption vector is introduced to characterize the resource usage of various types during a single inference process. For model i, its resource requirements can be expressed as:

$$c_i = [c_i^{cpu}, c_i^{gpu}, c_i^{mem}] \tag{3}$$

The system must satisfy the overall resource constraint at any given time step, meaning the sum of the resource consumption of all scheduled models must not exceed the system's available resource limit:

$$\sum_{i \in M_t} c_i \leq R_t \tag{4}$$

Here, M_t represents the set of models scheduled for execution at time step t, and R_t represents the vector of currently available resources in the system. This constraint ensures that resource contention will not cause system instability during the parallel execution of multiple models. Furthermore, by explicitly modeling resource constraints, the scheduling strategy can balance the collaborative execution of multiple models with resource utilization efficiency, thus providing a feasible decision space for adaptive scheduling.

3.3 Modeling of Adaptive Scheduling Decision Mechanism

To effectively respond to dynamic environments, adaptive scheduling mechanisms need to continuously adjust their decision-making behavior based on changes in system state. The scheduling process is modeled as a sequential decision problem, with the optimization objective being to minimize the long-term operating cost of the system. Therefore, an immediate system cost function is introduced to comprehensively characterize factors such as inference latency, resource consumption, and scheduling overhead:

$$l_t = \alpha \cdot D_t + \beta \cdot U_t \tag{5}$$

Where D_t represents the system delay metric at time step t, U_t represents the resource utilization-related cost, and α and β are trade-off coefficients. The overall system optimization objective is defined as minimizing the long-run cumulative cost:

$$\min_{\pi} \mathbb{E}\left[\sum_{t=0}^{T} l_t\right] \tag{6}$$

3.4 Cooperative optimization mechanism for scheduling and resource orchestration

In multi-model distributed inference systems, scheduling decisions and resource orchestration are highly coupled. To reflect their synergistic effect, a joint decision variable is introduced, unifying model selection and resource allocation into a single optimization process. For a given state s_t, the joint decision result can be expressed as:

$$a_t = (x_t, y_t) \tag{7}$$

Where x_t represents the model scheduling decision, and y_t represents the corresponding resource allocation scheme. To ensure the feasibility of system operation, the resource allocation result must meet the following constraints:

$$y_t = Y(x_t, R_t) \tag{8}$$

This collaborative optimization mechanism unifies the modeling, scheduling, and resource orchestration processes at the decision-making level, enabling the system to dynamically adjust its execution strategy based on model characteristics and resource status, thereby improving the overall operational efficiency and adaptability in multi-model collaborative reasoning scenarios.

4. Experimental Analysis

4.1 Dataset

This study adopts the Alibaba Cluster Trace Dataset, which is a publicly available large-scale system trace collected from a production-level computing cluster. The dataset records long-term operational information of a distributed environment that supports diverse computational workloads. It contains detailed logs of task

submission, execution behavior, and resource usage, which reflect realistic system dynamics under varying load conditions. Due to its scale and completeness, the dataset has been widely used to study scheduling, resource management, and system-level optimization problems.

The dataset provides fine-grained information on multiple workload types running concurrently in the cluster. Each workload is associated with heterogeneous resource requirements, including CPU usage, memory consumption, and execution duration. Tasks arrive over time with significant variability in arrival patterns and execution characteristics. This property makes the dataset well-suited for modeling multi-model collaborative inference scenarios, where different models exhibit distinct computational behaviors and compete for shared system resources.

In addition, the dataset captures dynamic changes in system state, such as resource availability fluctuations and workload intensity variations. These characteristics closely resemble real-world distributed inference environments, in which scheduling decisions must adapt to non-stationary conditions. By leveraging this open dataset, the study grounds the proposed scheduling framework in a realistic system context, enabling meaningful analysis of adaptive coordination mechanisms in multi-model distributed inference systems.

4.2 Experimental Results

This article first presents the results of the comparative experiments, as shown in Table 1.

Table 1. Comparative experimental results

Method	Resource Utilization	Load Balance	Stability	Adaptation Responsiveness	Method	Resource Utilization	Load Balance	Stability
iFLOW [8]	63.4	0.182	0.71	0.54	iFLOW [8]	63.4	0.182	0.71
Scar [9]	67.9	0.165	0.74	0.58	Scar [9]	67.9	0.165	0.74
HEART [10]	70.2	0.158	0.77	0.62	HEART [10]	70.2	0.158	0.77
Top [11]	72.5	0.149	0.80	0.66	Top [11]	72.5	0.149	0.80
MCCE [12]	74.1	0.142	0.82	0.69	MCCE [12]	74.1	0.142	0.82
Ours	79.6	0.118	0.89	0.78	Ours	79.6	0.118	0.89
Method	Resource Utilization	Load Balance	Stability	Adaptation Responsiveness	Method	Resource Utilization	Load Balance	Stability
iFLOW [8]	63.4	0.182	0.71	0.54	iFLOW [8]	63.4	0.182	0.71

The results indicate clear performance differences among the compared methods in multi-model collaborative environments. Adaptive scheduling strategies demonstrate more prominent and stable behavior in terms of resource utilization. Under identical system constraints, the adaptive approach exploits heterogeneous resources more effectively. It allows inference processes of different models to maintain higher resource occupancy in a shared environment. This capability is critical for the parallel execution of multiple models. It improves overall system processing capacity and reduces performance loss caused by resource underutilization.

Improvements brought by adaptive scheduling are particularly evident in load balancing. Compared with other methods, this strategy distributes computational pressure more evenly across computing nodes or model instances. Multi-model collaborative inference is often accompanied by differences in model scale and fluctuations in task intensity. If local hotspots are not mitigated in time, queue buildup and latency accumulation may occur. The results show that the adaptive strategy significantly alleviates such an imbalance. This provides a foundation for stable system operation under high load conditions.

From the perspective of system stability, adaptive scheduling exhibits smoother behavior during multi-model collaborative execution. It maintains lower performance variability under complex workload changes. Multi-model inference involves multiple computational paths running in parallel. Interference among models can amplify system-level instability. By continuously sensing system states and dynamically adjusting scheduling decisions, the adaptive strategy effectively reduces latency fluctuations. This enables more sustained stability during long-term operation.

Regarding responsiveness to workload variation, adaptive scheduling demonstrates higher agility. When request patterns or model invocation frequencies change, the system adjusts scheduling outputs within a short time. Inference execution quickly returns to a stable regime. This rapid response is essential for multi-model collaborative systems. Model switching, resource contention, and dynamic concurrency constantly perturb system states. The results show that the adaptive mechanism promptly corrects scheduling behavior under changing conditions. It maintains efficient system operation in dynamic environments and provides a more flexible solution for practical deployment.

To further investigate the role of scheduling hyperparameters in distributed inference systems, a systematic sensitivity analysis is conducted under different hyperparameter configurations. This analysis focuses on the impact of variations in a single scheduling control parameter on system behavior. It aims to characterize how scheduling decisions respond to parameter perturbations, and the experimental results are shown in Figure 2.

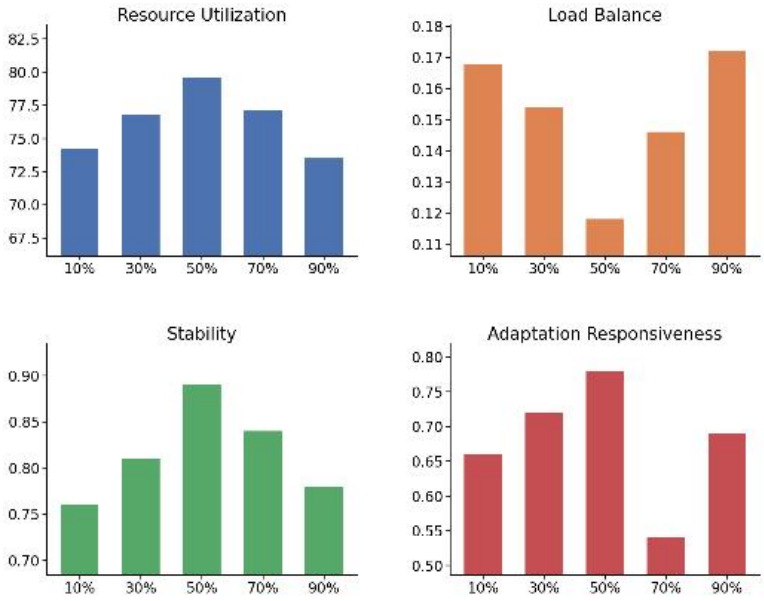

Figure 2. Experiments on the sensitivity of the system performance metrics with respect to variations in scheduling hyperparameters

The results indicate clear performance differences among the compared methods in multi-model collaborative environments. Adaptive scheduling strategies demonstrate more prominent and stable behavior in terms of resource utilization. Under identical system constraints, the adaptive approach

exploits heterogeneous resources more effectively. It allows inference processes of different models to maintain higher resource occupancy in a shared environment. This capability is critical for the parallel execution of multiple models. It improves overall system processing capacity and reduces performance loss caused by resource underutilization.

Improvements brought by adaptive scheduling are particularly evident in load balancing. Compared with other methods, this strategy distributes computational pressure more evenly across computing nodes or model instances. Multi-model collaborative inference is often accompanied by differences in model scale and fluctuations in task intensity. If local hotspots are not mitigated in time, queue buildup and latency accumulation may occur. The results show that the adaptive strategy significantly alleviates such an imbalance. This provides a foundation for stable system operation under high load conditions.

From the perspective of system stability, adaptive scheduling exhibits smoother behavior during multi-model collaborative execution. It maintains lower performance variability under complex workload changes. Multi-model inference involves multiple computational paths running in parallel. Interference among models can amplify system-level instability. By continuously sensing system states and dynamically adjusting scheduling decisions, the adaptive strategy effectively reduces latency fluctuations. This enables more sustained stability during long-term operation.

Regarding responsiveness to workload variation, adaptive scheduling demonstrates higher agility. When request patterns or model invocation frequencies change, the system adjusts scheduling outputs within a short time. Inference execution quickly returns to a stable regime. This rapid response is essential for multi-model collaborative systems. Model switching, resource contention, and dynamic concurrency constantly perturb system states. The results show that the adaptive mechanism promptly corrects scheduling behavior under changing conditions. It maintains efficient system operation in dynamic environments and provides a more flexible solution for practical deployment.

The scheduling update frequency is a critical control factor in distributed inference systems, as it affects decision timeliness and system response behavior. Different update frequencies change how the scheduling policy perceives variations in system states. This further influences the overall execution rhythm and resource allocation process. To characterize the behavior of scheduling mechanisms under parameter variations, a systematic sensitivity analysis of the scheduling update frequency at the system level is required, and the experimental results are shown in Table 2.

Table 2. System performance under varying scheduling update frequencies

Scheduling Update Frequency	Resource Utilization	Load Balance	Stability	Adaptation Responsiveness	Scheduling Update Frequency	Resource Utilization	Load Balance	Stability
0.2	72.8	0.176	0.74	0.63	0.2	72.8	0.176	0.74
0.5	75.9	0.162	0.78	0.67	0.5	75.9	0.162	0.78
1.0	79.6	0.118	0.89	0.78	1.0	79.6	0.118	0.89
2.0	77.2	0.154	0.82	0.72	2.0	77.2	0.154	0.82
4.0	73.4	0.169	0.76	0.65	4.0	73.4	0.169	0.76
Scheduling Update Frequency	Resource Utilization	Load Balance	Stability	Adaptation Responsiveness	Scheduling Update Frequency	Resource Utilization	Load Balance	Stability
0.2	72.8	0.176	0.74	0.63	0.2	72.8	0.176	0.74
0.5	75.9	0.162	0.78	0.67	0.5	75.9	0.162	0.78

From the overall trend, variations in the scheduling update frequency significantly affect system performance in multi-model collaborative scenarios. As the update frequency increases from a low level to a moderate range, the system can perceive state changes more promptly and adjust the inference process accordingly. Resource utilization improves steadily and reaches a higher level of effective activation. This indicates that a moderate scheduling pace strengthens cooperation among models and maintains high execution efficiency during parallel processing of complex tasks.

In terms of load balancing, the scheduling update frequency also has a pronounced impact on load distribution. At low frequencies, delayed scheduling signals make it easy for load disparities to emerge across computing nodes, which keeps the system in a relatively imbalanced state. When the update frequency reaches 1.0 times per second, the load balancing metric attains its optimal value. This shows that the scheduler can more effectively eliminate local hotspots and distribute computational pressure at this frequency. With excessively frequent updates, this capability weakens, indicating that scheduling overhead itself begins to interfere with system operation.

System stability exhibits a nonlinear pattern throughout the experiment and reaches its peak at a moderate frequency. When scheduling updates are insufficient, the system cannot promptly correct performance disturbances along the inference path, leading to a lower stability index. When the update frequency is too high, additional jitter is introduced, and the system experiences increased fluctuations due to continuous feedback changes. An intermediate update pace provides stronger self-regulation capability and enables the system to maintain smoother operation.

Regarding adaptive responsiveness, the moderate frequency also delivers the best performance. The system can rapidly make effective adjustments under dynamic load variations. Low-frequency scheduling suffers from delayed feedback and cannot respond in time when changes occur. High-frequency scheduling may cause conflicts due to frequent policy switching and reduce overall adjustment efficiency. These results indicate that adaptive scheduling with a reasonable update frequency preserves decision quality while maintaining agility. This allows multi-model collaborative inference to achieve stronger elasticity and controllability in dynamic environments.

Load fluctuation intensity is a key factor for characterizing environmental uncertainty in distributed inference systems. Its variation directly affects the effectiveness of scheduling decisions and the overall execution rhythm of the system. As request arrival patterns shift from stable to highly volatile, the system must preserve inference continuity and coordination under more complex environmental constraints. To examine the adaptability of scheduling mechanisms to such changes, a systematic environmental sensitivity analysis of system behavior under different levels of load fluctuation intensity is necessary, and the experimental results are shown in Figure 3.

From the overall trend, increasing load fluctuation intensity continuously raises environmental uncertainty. External disturbances faced by scheduling decisions are amplified accordingly. Under these conditions, resource utilization shows a pattern of gradual decline, partial recovery, and then a pronounced decrease. This reflects that the scheduling mechanism can maintain relatively high resource activation through dynamic adjustment under moderate fluctuations, while resource scheduling becomes significantly more difficult in high-intensity fluctuation environments.

The load balancing metric increases steadily as fluctuation intensity grows. This indicates that uneven distribution of computational pressure becomes more likely when environmental volatility intensifies. Although the scheduling strategy can effectively suppress load deviation under low to moderate fluctuations, frequent changes in request patterns combined with multi-model collaborative execution increase the cost of mitigating local hotspots. This imposes sustained pressure on overall scheduling efficiency.

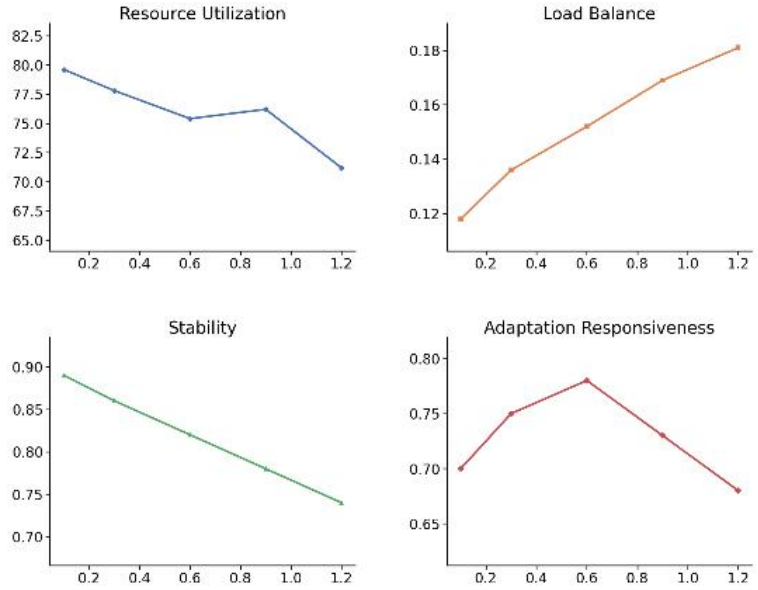

Figure 3. Sensitivity of system performance to varying load fluctuation intensity under the proposed adaptive scheduling mechanism in a multi-model collaborative inference environment

Regarding system stability, the stability index exhibits a smooth yet continuous decline as load fluctuation intensity increases. This shows that inference processes are influenced by the accumulation of multiple disturbances in dynamic environments. The scheduler must constantly balance response speed and execution consistency. Even so, the proposed method avoids severe oscillations and keeps performance variations within a controllable range, demonstrating strong robustness.

Changes in adaptive responsiveness show that the system exhibits the most proactive adjustment behavior under moderate load fluctuation intensity. It can quickly revise scheduling decisions to adapt to environmental changes. When fluctuations intensify further, the response magnitude declines, indicating that frequent and severe disturbances weaken the effectiveness of scheduling adjustments. Overall, these results demonstrate that the proposed adaptive scheduling mechanism maintains a reasonable response rhythm under complex load conditions and supports stable operation of multi-model collaborative inference systems in non-stationary environments.

The heterogeneity ratio of a cluster reflects the degree of variation in computing resources in terms of performance and architecture within a distributed inference system. Changes in this ratio directly affect the constraints and feasible space of scheduling decisions. In multi-model collaborative inference scenarios, shifts in the proportion of different node types reshape resource competition and task mapping patterns, which increases environmental complexity. To characterize the adaptability of scheduling mechanisms in heterogeneous settings, it is necessary to evaluate the environmental sensitivity of the system under different heterogeneity ratio conditions, and the experimental results are shown in Figure 4.

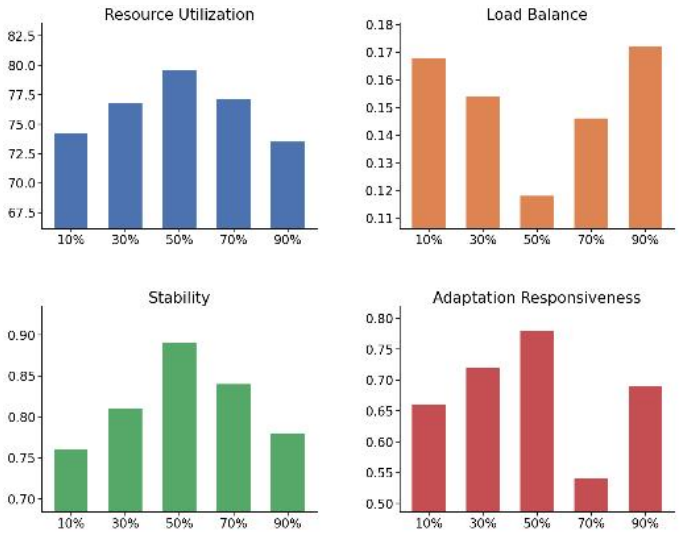

Figure 4. Impact of heterogeneous cluster ratio on system performance under the proposed adaptive scheduling framework in a multi-model collaborative inference setting

As the heterogeneity ratio changes, resource scheduling behavior in multi-model collaborative inference exhibits clear stage-dependent characteristics. When the proportion of heterogeneous nodes is low, the scheduling policy operates within a relatively homogeneous resource space. Decision flexibility is therefore constrained. As the heterogeneity ratio increases, the scheduler can exploit performance differences across nodes for more effective task mapping. This improves overall resource allocation and leads to a more favorable operating state at moderate heterogeneity levels.

From the perspective of load distribution, the degree of heterogeneity has a direct impact on computational balance within the system. When the heterogeneity ratio is either low or excessively high, performance differences among nodes cannot be fully absorbed. This often causes certain resources to become bottlenecks. The results show that under a moderate heterogeneity ratio, the scheduling strategy can more accurately match model characteristics with node capabilities. Computational pressure is thus distributed more reasonably across the cluster, which alleviates resource contention during multi-model parallel execution.

System stability varies with the heterogeneity ratio and shows an initial increase followed by a decline. This reflects the dual effect of heterogeneity on inference consistency. A moderate heterogeneous environment provides more feasible execution paths for scheduling decisions. This allows the system to maintain smoother operation when disturbances occur. When heterogeneity becomes excessive, the combined effects of performance disparity and communication overhead are amplified. This weakens the stability advantages of scheduling decisions.

In terms of adaptive responsiveness, the scheduling mechanism demonstrates more proactive adjustment behavior under moderate heterogeneity ratios. It can rapidly revise resource allocation to adapt to environmental changes. As heterogeneity increases further, the system must operate under more complex resource constraints. The effectiveness of scheduling adjustments is therefore reduced. Overall, the experiment indicates that the proposed adaptive scheduling method can fully leverage its strengths within a reasonable range of heterogeneous configurations. It supports the efficient operation of multi-model collaborative inference systems in complex cluster environments.

The model invocation ratio represents the relative frequency at which different inference models are activated and participate in execution within the system. Variations in this ratio directly shape the resource orientation of scheduling decisions and the selection of execution paths. In multi-model

collaborative inference scenarios, changes in invocation ratios are typically coupled with adjustments in computational load composition and resource competition patterns, which introduce stronger dynamics into system operation. Evaluating system behavior under varying model invocation ratios, therefore, requires a comprehensive data sensitivity analysis that reflects how scheduling mechanisms respond to shifts in model usage structure, and the experimental results are shown in Figure 5.

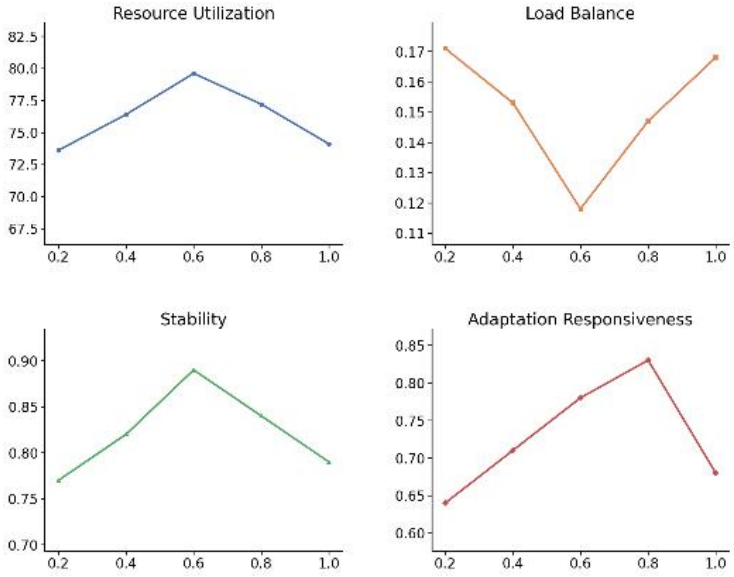

Figure 5. Sensitivity of system behavior to varying model invocation ratios under the proposed adaptive scheduling strategy in a multi-model collaborative inference system.

As the model invocation ratio gradually increases from low to high, the scheduling load structure in multi-model collaborative inference is continuously adjusted. Resource utilization reaches a relatively high level around moderate invocation ratios. This indicates that the scheduling mechanism can activate cluster resources more effectively when model participation remains balanced. When the invocation ratio increases further, certain models occupy more execution opportunities. Resource allocation flexibility is then constrained, which affects overall utilization efficiency.

With respect to load distribution, the model invocation ratio has a direct impact on computational balance within the system. When the ratio lies in a moderate range, the scheduling strategy distributes inference tasks more reasonably. Computational pressure across nodes is effectively alleviated. As the ratio shifts toward either extreme, model requests become more concentrated. Load distribution gradually becomes imbalanced, highlighting the restrictive role of model usage structure on scheduling performance in multi-model scenarios.

System stability exhibits clear nonlinear behavior as the model invocation ratio varies. Under moderate invocation conditions, interference among inference processes is well controlled. The system therefore operates in a smoother state. By contrast, when model invocation is overly concentrated or overly dispersed, the scheduler must frequently adjust execution order and resource mapping. This amplifies the risk of fluctuations during system operation.

Changes in adaptive responsiveness indicate that the scheduling mechanism shows more proactive adjustment behavior when the model invocation ratio is relatively high but not dominant. In this range, the system can respond quickly to shifts in model demand while preserving scheduling flexibility. When the ratio increases further, the response magnitude declines. This suggests that an imbalance in model usage reduces the adjustment space of adaptive scheduling. Overall, the results demonstrate that proper

control of the model invocation ratio is essential for maintaining efficiency and stability in multi-model inference systems.

5. Conclusion

This study focuses on scheduling in distributed inference systems under multi-model collaborative scenarios. It addresses the dual requirements of efficiency and stability for intelligent services in complex environments and establishes a unified analytical perspective on scheduling. As inference systems evolve from single model execution to parallel and collaborative multi-model operation, the role of scheduling becomes increasingly critical in resource coordination, execution organization, and system behavior control. The analysis shows that scheduling strategies with global awareness and dynamic decision making can better handle challenges arising from model diversity, resource heterogeneity, and dynamic request patterns, thereby offering an effective path to improving overall system performance.

From an application perspective, the findings provide valuable guidance for cloud-based inference services, edge computing platforms, and large-scale intelligent systems. In practical applications such as search, recommendation, perception, and analytics, multi-model collaboration has become a common paradigm, and scheduling strategies have a direct impact on service quality and resource cost. Looking ahead, continued growth in model scale and increasing diversity of inference scenarios will expose distributed inference systems to more complex operating conditions. Enhancing the ability of scheduling mechanisms to adapt to large-scale and highly diverse system states is likely to be a key direction for advancing intelligent inference infrastructure, and the present work offers a useful foundation and insight for this ongoing effort.

References

[1] A. Symons, L. Mei, S. Colleman, P. Houshmand, S. Karl and M. Verhelst, "Towards Heterogeneous Multi-Core Accelerators Exploiting Fine-Grained Scheduling of Layer-Fused Deep Neural Networks," arXiv preprint arXiv:2212, 2022.

[2] Z. Wang, "Federated Multi-Scale Representation Learning for Privacy-Aware Log Anomaly Detection in Distributed Cloud Environments," 2024.

[3] Z. Wang, Y. Yu, W. Zheng, W. Ma, and M. Zhang, "Macrec: A multi-agent collaboration framework for recommendation," in Proc. 47th Int. ACM SIGIR Conf. Research and Development in Information Retrieval, Jul. 2024, pp. 2760-2764.

[4] N. Williams, S. K. Suresh, L. Hughes, B. Kileen, and T. Galanos, "Applications of an LLM to scale and automate computational workflows for civil structural design," in Proc. IASS Annual Symposia, vol. 2024, no. 11, Aug. 2024, pp. 1-9.

[5] Q. Zhang, "Adaptive Resource Scheduling in Distributed Computing via Multi-Agent Reinforcement Learning and Graph Convolutional Modeling," 2024.

[6] Weerasooriya A, Wanniarachchi D, Peiris S H, et al. Multi-Model System for Sustainable Coral Reef Conservation in Sri Lanka[C]//2024 6th International Conference on Advancements in Computing (ICAC). IEEE, 2024: 504-509.

[7] Y. Ma, "Anomaly detection in microservice environments via conditional multiscale GANs and adaptive temporal autoencoders," 2024.

[8] Z. Cui, X. Xiao, W. Qiong, P. Fang, Q. Feng, H. Zhang, and J. Wang, "Anti-Byzantine attacks enabled vehicle selection for asynchronous federated learning in vehicular edge computing," China Communications, vol. 21, no. 8, pp. 1-17, 2024.

[9] Odema M, Chen L, Kwon H, et al. Scar: Scheduling multi-model ai workloads on heterogeneous multi-chiplet module accelerators[C]//2024 57th IEEE/ACM International Symposium on Microarchitecture (MICRO). IEEE, 2024: 565-579.

[10] Z. Qiu, "A Multi-Scale Deep Learning and Uncertainty Estimation Framework for Comprehensive Anomaly Detection in Cloud Environments," 2023.

[11] X. Yang, "Trend-Fluctuation Decomposition with Deep Residual Networks for System Forecasting," 2024.

[12] J. Wang, J. Wang, B. Athiwaratkun, C. Zhang, and J. Zou, "Mixture-of-agents enhances large language model capabilities," arXiv preprint arXiv:2406.04692, 2024.