2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC)
October 5-8, 2025. Vienna, Austria

MALE: A Multi-Objective Evaluation Method for AI Mobility Services across the Cloud-Edge-Device Continuum

Junhee Lee[1,2], Yong-Jun Shin[1], and SungJoo Kang[1]

***Abstract*— Deploying AI services on battery-powered mobility platforms such as autonomous vehicles, mobile robots, and large scale IoT sensor networks requires determining the most suitable execution environment for each workload across the cloud, edge, and device computing options. Because every placement option imposes different trade-offs among Accuracy, Latency, and Energy efficiency (ALE), stakeholders face a difficult, mission-critical decision that existing studies seldom address in a holistic, mission-aware manner. To fill this gap, we introduce the Mission-driven ALE (MALE) evaluation method. MALE couples ALE metrics with explicit mission objectives by allowing analysts to apply customizable weights to each criterion. The evaluation results are aggregated and visualized as heatmaps, helping transform a previously heuristic and opaque placement decision (black-box) into a more transparent and interpretable process (white-box). We examine the applicability of MALE through three representative case studies: Autonomous Vehicles, Real-Time Robotics, and IoT Sensor Networks, each reflecting distinct ALE priorities. By supplying a structured, mission-aware decision-support method, MALE strengthens stakeholder confidence and accelerates the optimization of AI service placement across the cloud–edge–device continuum, offering a practical foundation for future validation in real-world deployments.**

I. Introduction

The rapid advancement of artificial intelligence (AI) technologies is transforming various sectors, particularly mobility and robotics, by enabling sophisticated decision-making, real-time responsiveness, and efficient resource management. However, the deployment of AI-driven services in battery-dependent mobility devices such as autonomous vehicles, robots, and IoT sensor networks introduces unique challenges, as these systems must operate under strict constraints related to computational resources, latency, and energy consumption. Existing studies have extensively explored computing paradigms individually, such as cloud, fog, edge, and on-device computing, each offering distinct trade-offs in terms of Accuracy, Latency, and Energy efficiency (ALE) [1]–[3]. Although valuable, these studies typically evaluate these criteria in isolation or partially combined, without systematically addressing their integration with specific mission requirements [4]–[6]. This gap limits practical decision-making for stakeholders who require holistic evaluations to align computational architectures with precise operational goals.

[1] Artificial Intelligence Computing Research Laboratory, Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea {j.h.lee, yjshin, sjkang}@etri.re.kr
[2] School of Computing, KAIST, Daejeon, Republic of Korea the78910@kaist.ac.kr

To address this critical need, we propose the Mission-driven Accuracy, Latency, and Energy Efficiency (MALE) evaluation method. MALE uniquely integrates mission-specific objectives with comprehensive performance metrics to systematically assess and select optimal computing paradigms tailored explicitly to mobility-centric AI services. By providing a structured approach to evaluating computing paradigms, including cloud, edge, and device-based architectures, MALE enables stakeholders to balance the critical trade-offs among these criteria.

This paper contributes to existing literature by:

- Offering a comparative characterization of cloud, edge, and on-device computing paradigms for AI mobility services, grounded in ALE performance metrics.
- Introducing the MALE evaluation method using Likert-scale scoring and weighted assessment tailored to AI inference and training phases.
- Exploring the feasibility and practical relevance of the proposed method through three representative case studies, each reflecting distinct ALE priorities in industrial deployment scenarios.

Through this holistic approach, our method functions as a structured decision support system that assists decision-makers in exploring the decision space for deploying AI mobility services in alignment with their operational goals. As illustrated in Fig. 1, this approach replaces the opaque black box with a transparent and interpretable white box, enabling informed and mission-oriented decisions by using intuitive heatmap visualizations. The remainder of this paper is organized as follows: Section II reviews related works. Section III describes the MALE evaluation method in detail. Section IV presents industrial case studies demonstrating the method's applicability. Section V discusses key limitations, and future works. Finally, Section VI concludes the paper.

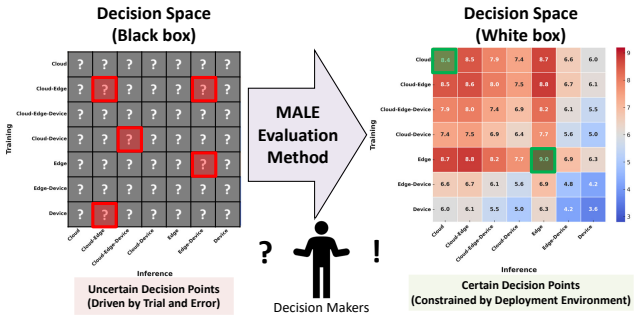

Fig. 1. The proposed method contributes by converting the decision space into a white-box model, clarifying uncertain decision points.

II. Related Works

Previous studies have explored cloud, fog, edge, and on-device computing paradigms by evaluating various performance metrics, such as computational capacity, latency, and energy consumption [7]–[9]. For instance, Zhou et al. [10] proposed a six-level rating system specifically designed for edge intelligence, which provided structured insights into distributed architectures. However, these studies mainly focused on specific technical characteristics without fully integrating mission-specific requirements. Several studies have individually highlighted key criteria:

- **Accuracy:** Highlighted by studies such as [6], emphasizing the significance of accuracy for reliable decision-making, particularly in safety-critical contexts.
- **Latency:** Examined extensively in real-time applications, such as autonomous driving and surveillance, where latency impacts system safety and performance [11], [12].
- **Energy Efficiency:** Investigated for its critical role in enhancing operational longevity for battery-dependent mobility devices, with several proposed methods to reduce energy usage [4], [13].

To better understand the extent and limitations of existing work, Table I provides a structured overview of recent studies, highlighting their coverage of cloud, edge, and device computing environments along with their assessment of ALE and mission-specific alignment. From this comparison, we identify significant observations:

- Most studies address only one or two performance criteria (ALE), but rarely all three simultaneously.
- Mission-specific considerations are rarely addressed or only loosely integrated, highlighting a gap between general performance metrics and their contextual application to real-world deployment scenarios.
- Limited research exists on methods explicitly integrating all three performance metrics (ALE) with clearly defined mission objectives, especially within the context of AI mobility services.

TABLE I
Comparison of Research Papers on MALE Metrics in Cloud, Edge, and Device Computing Environments

References	Cloud	Edge	Device	Mission	Accuracy	Latency	Energy
Ko et al. [1]	✓	✓	×	×	✓	✓	✓
Lockhart et al. [2]	✓	✓	✓	×	×	✓	×
Hu et al. [3]	✓	✓	✓	×	✓	✓	✓
Teerapittayanon et al. [5]	✓	✓	✓	×	✓	✓	×
Kang et al. [14]	✓	×	✓	×	×	✓	✓
Han et al. [15]	✓	×	✓	×	✓	✓	✓
Eshratifar et al. [16]	✓	×	✓	×	×	✓	✓
Li et al. [17]	✓	×	✓	×	✓	✓	×
Jeong et al. [18]	×	✓	✓	×	×	✓	✓
Proposed Method	✓	✓	✓	✓	✓	✓	✓

Our research aims to address these gaps by proposing the MALE evaluation method, which systematically integrates mission-specific requirements with the critical metrics of ALE. This approach provides a more holistic and structured framework for evaluating and deploying AI services in battery-dependent mobility contexts.

III. MALE Evaluation Method

In this section, we introduce the MALE evaluation method specifically designed for AI service deployment in battery-dependent mobility devices. The method systematically integrates mission-specific objectives with three critical performance metrics: **Accuracy**, **Latency**, and **Energy Efficiency**. By enabling customizable weighting of ALE metrics based on application-specific requirements, MALE provides a structured framework for assessing and comparing diverse computing paradigms. This allows stakeholders to visualize trade-offs among competing performance goals and make informed deployment decisions under practical constraints. As AI services increasingly operate on resource-limited platforms, such as mobile robots and embedded devices, this method supports balanced optimization of performance, responsiveness, and energy consumption tailored to each deployment context.

A. Characterizing AI Mobility Services within the Edge-Cloud Continuum

AI-based mobility services can be deployed across various computing paradigms, each offering unique advantages and trade-offs concerning ALE metrics. This structured characterization provides a foundation for further discussion and analysis within the MALE evaluation method, guiding deployment decisions tailored explicitly to mission-specific ALE requirements.

1) Defining mission based on ALE: The optimal configuration of computing architectures significantly depends on mission-specific priorities, specifically determined by the relative importance assigned to ALE. Unlike the fixed weight combinations utilized in previous analyses, our proposed approach emphasizes a systematic method for defining ALE weights tailored to diverse mission requirements.

In real-world deployments, each mission possesses distinct operational constraints and performance goals, necessitating flexible weight assignment for ALE metrics. For instance:

- **Accuracy-critical missions** (e.g., medical diagnostics, autonomous vehicles): Accuracy receives primary emphasis to guarantee precise outcomes, while Latency and Energy Efficiency weights are assigned according to practical resource availability and response-time tolerance.
- **Latency-critical missions** (e.g., real-time robotics): Latency assumes the highest weight to ensure prompt decision-making, with secondary importance placed on Accuracy or Energy Efficiency based on the context.
- **Energy-sensitive missions** (e.g., IoT sensors, battery-powered devices): High priority is given to Energy Efficiency, followed by either Latency or Accuracy, depending on whether timely data or precise analytics is more critical.

TABLE II
COMPARATIVE ANALYSIS OF COMPUTING PARADIGMS RELEVANT TO AI TRAINING AND INFERENCE [19]–[25]

Aspect	Cloud Computing	Fog Computing	Edge Computing	Multi-Access Edge Computing (MEC)	On-Device Computing
Definition	Centralized computing with virtually unlimited resources in large data centers.	Distributed computing via nearby intermediate nodes extending cloud services.	Local processing near data sources to reduce central dependency.	Computing at mobile network base stations close to users.	Processing directly on user devices without external reliance.
Computing Capacity	Very High (virtually unlimited scalable resources).	Moderate to High (regional nodes with moderate computing power).	Moderate (local servers capable of moderate complexity AI).	Moderate to High (powerful localized resources at cell sites).	Very Low (limited by device hardware and power constraints).
Latency	High (network delay due to distance from users, tens to hundreds of ms).	Moderate to Low (few to tens of ms; shorter distances reduce delay).	Very Low (local LAN processing, few ms latency).	Ultra-Low (1–10 ms, optimized within mobile network proximity).	Minimal (virtually zero network delay, instantaneous response).
Device Energy Consumption	Low (offloads processing to the cloud, minimizing device energy use).	Lower–Moderate (local nodes reduce transmission energy and processing load).	Low (offloads computation nearby, reducing device power usage).	Moderate (close processing reduces latency but increases communication overhead).	High (device performs all computation, draining battery directly).
AI Inference Suitability	Excellent for complex inference, allowing very large models.	Good for localized IoT inference, moderate-sized models.	Highly suitable for real-time inference using moderately complex models.	Highly suitable for mobile low-latency inference tasks.	Limited to lightweight inference tasks due to resource constraints.
AI Training Suitability	Outstanding (ideal for large-scale model training and complex deep learning).	Limited (small-scale or incremental updates, no heavy training).	Moderate (small-scale model training or local fine-tuning).	Moderate (localized federated learning, small-scale training).	Very Poor (only minimal training feasible, e.g., federated learning or personalization).

Our method provides decision-makers with an adaptable weighting method, allowing them to define ALE priorities explicitly aligned with their mission contexts. By systematically adjusting these weights and evaluating corresponding heatmaps, stakeholders can comprehensively assess various architectural combinations and select the suitable solution for their specific operational needs. This significantly extends the flexibility and applicability of performance evaluation compared to the black-box method, highlighting the importance of customizable ALE prioritization.

2) Classifying the service instances in the continuum: To support the classification of service instances within the edge–cloud continuum, we provide a structured summary and comparison of key attributes across widely-adopted computing paradigms, as shown in Table II. This comparative analysis outlines essential characteristics such as computing capacity, latency, proximity to users, device energy consumption, and security considerations. It also relates these characteristics to typical use cases in AI mobility services, thereby helping analysts make more informed architectural decisions. By mapping AI mobility services to this classification space, stakeholders can determine suitable deployment strategies for the specific ALE metrics relevant to the missions.

B. MALE Evaluation of the AI Mobility Service in Edge-Cloud Continuum

1) Setting MALE Score Criteria: This subsection describes the criteria used to establish MALE scores clearly and systematically. The evaluation criteria are based on ALE metrics selected to represent critical performance dimensions for AI mobility services within edge-cloud environments:

- **Accuracy**: The achievable model accuracy under average computational capabilities across different computing paradigms for identical tasks and constraints.
- **Latency**: Response time experienced at the device level.
- **Energy Efficiency**: Energy consumption of the device during computational activities.

Each computing paradigm is scored from 1 (poor) to 5 (excellent) for the training and inference phases, based on extensive literature reviews and empirical benchmarks, as shown in Fig. 2. Seven radar graphs represent all possible combinations of Cloud, Edge, and Device deployments—three single, three dual, and one full integration. Tables II and III detail the rationale behind each Likert-scale score assignment. These rationales ensure transparency, accuracy, and reproducibility, and are derived through comprehensive comparative analyses of previous studies [19]–[23].

2) Calculating MALE Score: This subsection emphasizes the quantitative calculation of the MALE score, which integrates predefined criteria with mission-specific weights to facilitate interpretable, white-box visualization through heatmap graphs. The total MALE score for each architecture is calculated by assigning user-defined numerical weights to each of the ALE criteria. The generalized formula for computing the MALE score is defined as follows:

$$\text{MALE\_Score} = \sum_{c \in \mathcal{C}} w_c \cdot (c_{\text{train}} + c_{\text{infer}}) \tag{1}$$

where $\mathcal{C} = \{\text{Accuracy}, \text{Latency}, \text{Energy}\}$, $\sum_{c \in \mathcal{C}} w_c = 1$, and $c_{\text{train}}, c_{\text{infer}} \in \{1, \ldots, 5\}$ for all $c \in \mathcal{C}$.

These weights allow analysts to explicitly prioritize the ALE criteria based on the mission-specific objectives. Adjusting these weights provides the flexibility to conduct sensitivity analyses and iterative recalculations, resulting in multiple white-box heatmap visualizations. These heatmaps effectively illustrate trade-offs across ALE dimensions, thereby enabling informed and transparent decision-making regarding architecture selection for AI mobility services. In sum-

TABLE III
RATIONALE FOR LIKERT-SCALE SCORES (1–5) OF COMPUTING PARADIGMS [19]–[26]

Criterion	Paradigm	Score	Rationale
Accuracy (Training)	Cloud	5	Cloud computing supports training of highly complex models with large datasets due to abundant computing and storage resources.
	Edge	4	Edge servers offer moderate compute capabilities for training, allowing for some model complexity but not to the scale of cloud environments.
	On-Device	1	On-device training is extremely constrained due to low compute power and limited memory, only supporting lightweight models with reduced accuracy.
	Hybrid (Fog/MEC)	2–4	Fog and MEC nodes can participate in distributed or federated learning, reaching near-cloud accuracy depending on aggregation strategy.
Accuracy (Inference)	Cloud	5	Cloud-hosted models can be full-scale and state-of-the-art, enabling the highest possible inference accuracy.
	Edge	4	Edge inference often runs moderately complex models that trade off some accuracy for latency and energy savings.
	On-Device	1	On-device models are typically quantized or pruned to fit local hardware constraints, resulting in modestly reduced accuracy.
	Hybrid (Fog/MEC)	2–4	Inference can be partitioned between fog and cloud, achieving both high speed and cloud-level accuracy via model splitting and layered refinement.
Latency (Training)	Cloud	1	High latency due to remote location and network delays, even though computation itself is fast in the cloud.
	Edge	3	Edge training reduces data upload time but operates under lower compute power, leading to moderate training latency.
	On-Device	5	Training happens locally without communication delays; however, actual training time may still be long.
	Hybrid (Fog/MEC)	2–4	Latency varies depending on node proximity and workload distribution, ranging from moderate (fog) to low (MEC).
Latency (Inference)	Cloud	1	Inference over the cloud involves round-trip network latency and potential congestion, often unacceptable for real-time requirements.
	Edge	3	Edge inference reduces latency but may still suffer from moderate delay due to network stack and service orchestration.
	On-Device	5	On-device inference achieves near-instant responses with no network overhead, ideal for ultra-low-latency needs.
	Hybrid (Fog/MEC)	2–4	Latency ranges from moderate (fog) to ultra-low (MEC), depending on deployment specifics and workload partitioning.
Energy Efficiency (Training)	Cloud	5	Training is offloaded, so device energy is preserved, though data transmission consumes some energy.
	Edge	5	Short-range transmission and remote computation ensure minimal device-side energy consumption.
	On-Device	1	Training consumes a large amount of energy on the device, which may overheat or rapidly deplete battery.
	Hybrid (Fog/MEC)	2–5	Depending on proximity and workload division, energy usage ranges from moderate (fog) to highly efficient (MEC).
Energy Efficiency (Inference)	Cloud	5	Cloud inference avoids computation on the device, saving energy, but incurs some cost from wide-area transmission.
	Edge	5	Nearby edge inference avoids computation on the device and minimizes transmission distance, preserving battery.
	On-Device	1	Inference computation happens locally, using device resources and potentially draining battery during continuous operation.
	Hybrid (Fog/MEC)	2–5	Energy usage varies by configuration: fog scenarios may require moderate energy, while MEC provides efficient edge processing.

mary, the MALE evaluation method significantly enhances architectural selection by systematically quantifying key performance trade-offs, thus providing a robust, transparent, and mission-aligned framework to optimize the deployment of AI mobility services within cloud-edge-device environments.

IV. CASE STUDIES

A. Three Industrial Cases Under Evaluation

This section provides a detailed analysis of ALE weight assignments informed by real-world implementations, technological reports, and industry literature. While the weights are not derived directly from raw field data, they are grounded in patterns and priorities consistently observed across documented use cases and expert sources.

1) Autonomous Vehicles (Accuracy Prioritized): Autonomous vehicles, such as Tesla Autopilot, require high accuracy due to stringent safety and reliability constraints [27]. High accuracy minimizes misclassification risks that directly contribute to accidents. Latency is critically important to ensure timely responses (milliseconds scale). Energy efficiency remains relevant but is typically secondary, as onboard power sources are sufficiently robust.

2) Real-Time Robotics (Latency Prioritized): Latency minimization is paramount in real-time robotics applications, exemplified by Boston Dynamics' Spot, due to immediate feedback requirements for precise control [28]. Literature and product documentation clearly indicate latency as the highest priority. Energy efficiency is also significant, particularly due to battery-operated constraints. Accuracy, while still essential, can be managed to acceptable levels without compromising overall real-time functionality.

3) IoT Sensor Networks (Energy Efficiency Prioritized): In IoT sensor networks such as LoRaWAN and Amazon

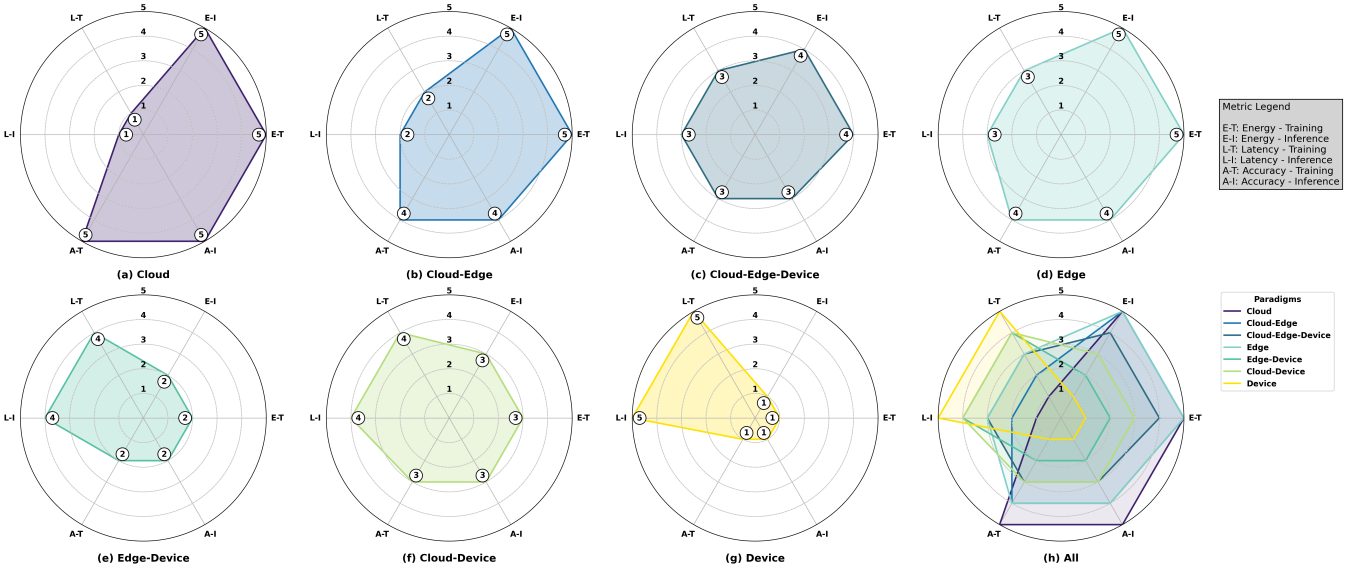

Fig. 2. Radar charts visualizing Likert-scale (1–5) performance scores across six evaluation criteria for each computing paradigm.

Sidewalk, energy efficiency is the concern due to limited battery power and long-term deployment requirements [29], [30]. Industry use-cases demonstrate greater tolerance for latency than inaccuracies or data reliability issues. Energy efficiency is therefore prioritized, with accuracy maintaining importance over latency to ensure data integrity and reliability. In summary, these scenarios illustrate distinct ALE priority trade-offs, providing practical insights into system design considerations for varying application contexts.

B. Evaluation Results

This section evaluates the computing paradigm selection method by applying six distinct ALE weight configurations, each representing distinct prioritizations of ALE. These configurations are derived from industrial requirements, product implementations, and additional hypothetical scenarios for broader coverage. Fig. 3 shows six corresponding heatmaps, each illustrating the total weighted MALE score (computed using Equation 1) across various architectural combinations. Among these six configurations, three closely match the industrial scenarios detailed in Section IV. These three (Autonomous Vehicles, Real-Time Robotics, IoT Sensor Networks) are discussed in detail because they exemplify the most prominent, validated use cases:

- **Autonomous Vehicles: Accuracy Prioritized**
  - **Weights:** Accuracy = 0.6, Latency = 0.3, Energy = 0.1
  - Autonomous systems such as Tesla Autopilot require highly accurate perception for safety, while also demanding rapid inference for real-time responsiveness [31]. Fig. 3(a) confirms that high-accuracy and low-latency-capable architectures such as **Cloud** and **Edge** platforms yield the highest total MALE scores under this revised weighting. However, since **Cloud-Device** configurations also achieve competitive scores, this suggests that selecting the absolute top-scoring option is not always necessary—especially in emergency scenarios where immediate inference is crucial, performing inference directly on the **Device** may offer practical advantages over relying on Cloud or Edge computation.
- **Real-Time Robotics: Latency Prioritized**
  - **Weights:** Accuracy = 0.1, Latency = 0.7, Energy = 0.2
  - Real-time robotic systems, such as Boston Dynamics' Spot, require ultra-low latency to ensure stable control and safe motion execution. For instance, Spot is designed to accept and execute control commands within a 300 ms time span; exceeding this latency can noticeably affect performance [32].
  - Fig. 3(c) indicates that **Device**, **Edge Edge-Device**, and **Cloud-Device** platforms perform best under these latency-focused requirements.
- **IoT Sensor Networks: Energy Prioritized**
  - **Weights:** Accuracy = 0.2, Latency = 0.1, Energy = 0.7
  - In long-term IoT deployments such as LoRaWAN and Amazon Sidewalk, power consumption is the primary constraint due to limited battery resources [29], [33]. Fig. 3(e) demonstrates that **Edge** and **Cloud-Edge** architectures provide optimal energy efficiency while preserving acceptable accuracy and latency.

The additional three weight configurations (depicted in Fig. 3(b), (d), and (f)) explore further hypothetical scenarios to broaden the design space analysis. While not elaborated upon here, they provide supplementary insights into potential trade-offs under different priority distributions. These six heatmaps confirm that the proposed method can flexibly align architectural decisions with a variety of application constraints by integrating interpretable, priority-driven evaluations. The three highlighted examples illustrate how ALE-related domains can benefit from architectural configurations to optimize critical performance metrics.

V. Discussion

A. Threats to Validity

Several factors may influence the validity of our study. Firstly, the subjective assignment of scores to accuracy,

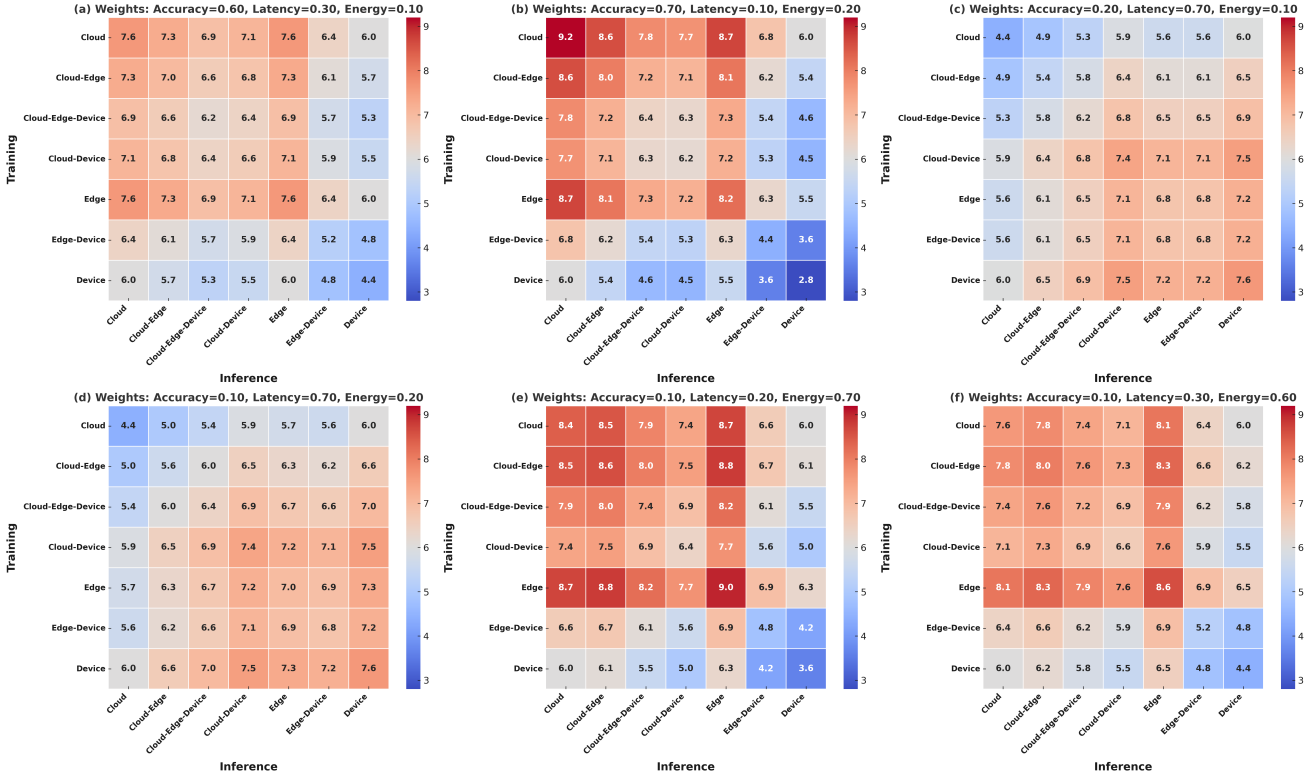

Fig. 3. Heatmap of total weighted scores for various architectural configurations under different mission-specific weight settings.

latency, and energy efficiency metrics could introduce biases. However, our study establishes a baseline reflecting typical scenarios by referencing established literature and industry-standard benchmarks. [19]–[25] This baseline is explicitly designed to serve as a foundational reference, enabling stakeholders to transparently adjust and customize the scoring criteria according to their specific application needs or priorities. To further mitigate potential biases, we provided detailed scoring methods.

Secondly, the selected case studies—autonomous vehicles, real-time robotics, and IoT sensor networks—were primarily chosen to demonstrate the applicability of our proposed method. While the validity of individual MALE analysis results for each case was not strictly within our study's scope, we have attempted to ensure accuracy and relevance by consulting multiple sources from existing literature and industry information [27]–[31], [33]. Nevertheless, generalizing our findings to other domains may require additional scenario-specific considerations and validation efforts. Future studies could address this by extending evaluations to a broader range of applications and industrial contexts.

Lastly, our evaluation method assumes static mission objectives within each scenario. Real-world operations often involve dynamic and evolving mission requirements, which could necessitate continuous recalibration of ALE priorities. Future work should incorporate adaptive mechanisms to dynamically adjust criteria weights in response to real-time operational changes.

B. Value of Our Method and Future Works

The MALE evaluation method presented in this study significantly advances current practices by integrating mission-specific objectives explicitly with ALE metrics. Unlike previous studies addressing isolated performance criteria, our comprehensive approach enables decision-makers to systematically assess computing paradigms tailored to their unique operational needs.

The practical case studies demonstrate the applicability and effectiveness of our method across diverse industrial contexts: Autonomous Vehicles, Real-Time Robotics, IoT Sensor Networks, providing valuable insights for deploying AI services in battery-dependent mobility environments. This structured and adaptable method offers stakeholders a robust tool for optimizing their deployment strategies, balancing mission-critical priorities effectively.

In future work, we plan to apply the proposed method to real-world deployment environments. By evaluating its effectiveness in practical AI mobility systems, we aim to assess its potential as a foundational tool for optimizing deployment strategies across the cloud–edge–device continuum.

VI. Conclusion

We proposed the MALE (Mission-driven Accuracy, Latency, and Energy Efficiency) evaluation method to systematically guide the deployment of AI services for battery-dependent mobility devices across the cloud-edge-device continuum. Recognizing the limitations of existing works that evaluate performance metrics such as ALE in isolation or without explicit mission context, our approach introduces a structured and adaptable framework that aligns computing paradigm selection with mission-specific objectives.

This study's core contribution is the transformation of architectural decision-making from a heuristic and black-

box process into an interpretable and white-box method. By integrating customizable ALE priorities and visualizing trade-offs through heatmaps, the MALE method enables stakeholders to make informed, transparent, and mission-aligned deployment decisions across cloud, edge, and device computing paradigms. Through three representative case studies—Autonomous Vehicles, Real-Time Robotics, and IoT Sensor Networks—we examined how the proposed method can reflect diverse operational priorities. These examples indicate the potential of the MALE method to assist in balancing performance trade-offs in mission-dependent deployments. In future work, we plan to apply the proposed method to real-world deployment environments. By evaluating its effectiveness in practical AI mobility systems, we aim to assess its potential as a foundational tool for optimizing deployment strategies across the cloud–edge–device continuum.

ACKNOWLEDGMENT

This work was supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No.2024-00406245, Development of Software-Defined Infrastructure Technologies for Future Mobility)

REFERENCES

[1] J. H. Ko, T. Na, M. F. Amir, and S. Mukhopadhyay, "Edge-host partitioning of deep neural networks with feature space encoding for resource-constrained internet-of-things platforms," in *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–6, IEEE, 2018.

[2] L. Lockhart, P. Harvey, P. Imai, P. Willis, and B. Varghese, "Scission: Performance-driven and context-aware cloud-edge distribution of deep neural networks," in *2020 IEEE/ACM 13th International Conference on Utility and Cloud Computing (UCC)*, pp. 257–268, IEEE, 2020.

[3] S. Hu, C. Dong, and W. Wen, "Enable pipeline processing of dnn co-inference tasks in the mobile-edge cloud," in *2021 IEEE 6th International Conference on Computer and Communication Systems (ICCCS)*, pp. 186–192, IEEE, 2021.

[4] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," *arXiv preprint arXiv:1510.00149*, 2015.

[5] S. Teerapittayanon, B. McDanel, and H.-T. Kung, "Distributed deep neural networks over the cloud, the edge and end devices," in *2017 IEEE 37th international conference on distributed computing systems (ICDCS)*, pp. 328–339, IEEE, 2017.

[6] A. Vega, A. Buyuktosunoglu, D. Callegaro, M. Levorato, and P. Bose, "Cloud-backed mobile cognition: Power-efficient deep learning in the autonomous vehicle era," *Computing*, pp. 1–19, 2022.

[7] C. Gong, F. Lin, X. Gong, and Y. Lu, "Intelligent cooperative edge computing in internet of things," *IEEE Internet of Things Journal*, vol. 7, no. 10, pp. 9372–9382, 2020.

[8] J.-H. Huh and Y.-S. Seo, "Understanding edge computing: Engineering evolution with artificial intelligence," *IEEE Access*, vol. 7, pp. 164229–164245, 2019.

[9] S. Duan, D. Wang, J. Ren, F. Lyu, Y. Zhang, H. Wu, and X. Shen, "Distributed artificial intelligence empowered by end-edge-cloud computing: A survey," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 1, pp. 591–624, 2022.

[10] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, "Edge intelligence: Paving the last mile of artificial intelligence with edge computing," *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1738–1762, 2019.

[11] Y. Liu, Y. Deng, A. Nallanathan, and J. Yuan, "Machine learning for 6g enhanced ultra-reliable and low-latency services," *IEEE Wireless Communications*, vol. 30, no. 2, pp. 48–54, 2023.

[12] M. Satyanarayanan, "The emergence of edge computing," *Computer*, vol. 50, no. 1, pp. 30–39, 2017.

[13] A. Mughees, M. Tahir, M. A. Sheikh, and A. Ahad, "Towards energy efficient 5g networks using machine learning: Taxonomy, research challenges, and future research directions," *Ieee Access*, vol. 8, pp. 187498–187522, 2020.

[14] Y. Kang, J. Hauswald, C. Gao, A. Rovinski, T. Mudge, J. Mars, and L. Tang, "Neurosurgeon: Collaborative intelligence between the cloud and mobile edge," *ACM SIGARCH Computer Architecture News*, vol. 45, no. 1, pp. 615–629, 2017.

[15] S. Han, H. Shen, M. Philipose, S. Agarwal, A. Wolman, and A. Krishnamurthy, "Mcdnn: An approximation-based execution framework for deep stream processing under resource constraints," in *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*, pp. 123–136, 2016.

[16] A. E. Eshratifar, M. S. Abrishami, and M. Pedram, "Jointdnn: An efficient training and inference engine for intelligent mobile cloud computing services," *IEEE transactions on mobile computing*, vol. 20, no. 2, pp. 565–576, 2019.

[17] H. Li, C. Hu, J. Jiang, Z. Wang, Y. Wen, and W. Zhu, "Jalad: Joint accuracy-and latency-aware deep structure decoupling for edge-cloud execution," in *2018 IEEE 24th international conference on parallel and distributed systems (ICPADS)*, pp. 671–678, IEEE, 2018.

[18] H.-J. Jeong, H.-J. Lee, C. H. Shin, and S.-M. Moon, "Ionn: Incremental offloading of neural network computations from mobile devices to edge servers," in *Proceedings of the ACM symposium on cloud computing*, pp. 401–411, 2018.

[19] S. S. Gill, P. Garraghan, and R. Buyya, "Router: Fog enabled cloud based intelligent resource management approach for smart home iot devices," *Journal of Systems and Software*, vol. 154, pp. 125–138, 2019.

[20] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog computing and its role in the internet of things," pp. 13–16, 2012.

[21] W. Shi and S. Dustdar, "The promise of edge computing," *Computer*, vol. 49, no. 5, pp. 78–81, 2016.

[22] X. Wang, Y. Han, V. C. Leung, D. Niyato, X. Yan, and X. Chen, "Convergence of edge computing and deep learning: A comprehensive survey," *IEEE communications surveys & tutorials*, vol. 22, no. 2, pp. 869–904, 2020.

[23] L. Capogrosso, F. Cunico, D. S. Cheng, F. Fummi, and M. Cristani, "A machine learning-oriented survey on tiny machine learning," *IEEE Access*, vol. 12, pp. 23406–23426, 2024.

[24] P. Cong, J. Zhou, L. Li, K. Cao, T. Wei, and K. Li, "A survey of hierarchical energy optimization for mobile edge computing: A perspective from end devices to the cloud," *ACM Computing Surveys (CSUR)*, vol. 53, no. 2, pp. 1–44, 2020.

[25] Y. Long, I. Chakraborty, G. Srinivasan, and K. Roy, "Complexity-aware adaptive training and inference for edge-cloud distributed ai systems," in *2021 IEEE 41st International Conference on Distributed Computing Systems (ICDCS)*, pp. 573–583, IEEE, 2021.

[26] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE internet of things journal*, vol. 3, no. 5, pp. 637–646, 2016.

[27] Tesla, "Autopilot and full self-driving capability." `https://www.tesla.com/autopilot`. Accessed: 2025-04-18.

[28] B. Dynamics, "Spot sdk documentation." `https://dev.bostondynamics.com`. Accessed: 2025-04-18.

[29] Semtech, "Lorawan specification and architecture." `https://lora-alliance.org/resource_hub/what-is-lorawan`. Accessed: 2025-04-18.

[30] Amazon, "Introducing amazon sidewalk." `https://developer.amazon.com/en-US/blogs/alexa/device-makers/2020/09/amazon-sidewalk-paves-the-way-for-more-connected-communities`. Accessed: 2025-04-18.

[31] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, "A survey of deep learning techniques for autonomous driving," *Journal of field robotics*, vol. 37, no. 3, pp. 362–386, 2020.

[32] Boston Dynamics, "Prepare your network for spot." `https://support.bostondynamics.com/s/article/Prepare-Your-Network-for-Spot-49942`, 2023. Accessed: 2025-04-18.

[33] X. Li, F. Yu, and H. Zhang, "An energy-efficient framework for iot communication in smart homes," *IEEE Access*, vol. 7, pp. 150703–150713, 2019.