An Intent-Based Management Framework for On-Device Artificial Intelligence in Smart Factory

Yoseop Ahn*, Mose Gu*, and Jaehoon (Paul) Jeong*
* Department of Computer Science & Engineering,
Sungkyunkwan University, Suwon, Republic of Korea
Email:{ahnjs124, rna0415, pauljeong}@skku.edu

***Abstract*—This paper presents an Intent-Based On-Device Artificial Intelligence (AI) Framework that enables collaborative AI functionality among devices by reflecting user's intent. The proposed framework leverages multimodal sensor data and gives a policy to the devices through components such as a Speech-To-Text (STT) Model, Intent Translator, and Policy Coordinator. It runs On-Device AI through collaboration between individual devices, minimizes dependence on cloud-based systems, and considers information security, increased processing speed, and energy efficiency. The framework's applicability is explained through a use case involving emergency response scenarios, highlighting its potential in managing critical situations through intelligent coordination among Internet-of-Things (IoT) devices. Future work includes optimizing network intent integration and developing testbeds for performance evaluation in diverse IoT environments.**

***Index Terms*—Intent, On-Device AI, IoT Edge, Multimodal Sensor, Emergency Response Systems**

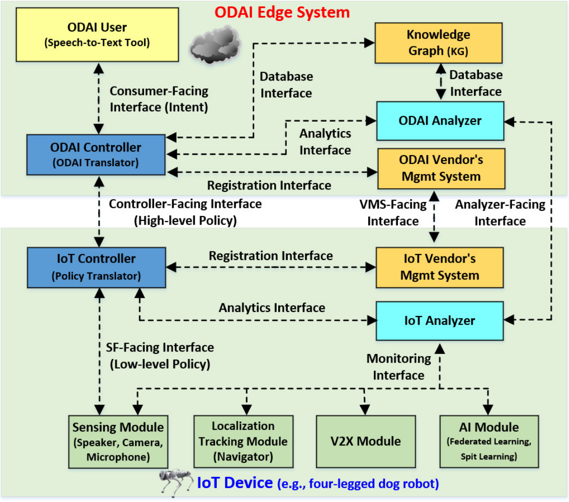

Fig. 1: Components and Interfaces of Intent-Based Management Framework for On-Device AI

I. Introduction

People began to have an interest in AI when AlphaGo appeared in 2016. Then, with the appearance of ChatGPT [1] in 2022, generative AI began to spread, and at the same time, as AI experiences became more popular among the public, the influence of AI in everyday life became very large. In particular, the core of a future IoT technology based on devices such as smartphones is to analyze, predict, and judge situations on their own to provide intelligent services with users, so the role of an object-centered AI is growing in the development of intelligent autonomous IoT, and an interest in On-Device AI used within devices is also continuously growing. On-device AI [2] refers to technology that runs AI algorithms and models on the device itself without connecting to a cloud server. In particular, due to issues such as the delay through network infrastructure that may arise when utilizing existing cloud-based AI, concerns about sensitive information leakage, and additional delay due to network load, the demand for On-Device AI functions that enable real-time information processing on personal devices such as smartphones without relying on the cloud is steadily increasing [3]. Also, On-Device AI can be used to complement the limitations of cloud computing, such as strengthening information security, improving service speed, and reducing energy consumption. It is also used to develop a deep learning technology and a low-power AI technology. However, there are still limitations in autonomously judging situations based on a single device's AI due to the lack of intelligence and computational capabilities of the device.

We propose an On-Device AI (ODAI) framework that solves AI problems through device collaboration based on a user's intent. This framework focuses on enabling devices to perform an optimal collaboration and a behavior by reflecting human intentions. Based on individual multimodal sensor data [4]. This ODAI framework enables them to communicate and cooperate with each other through an intent and forms an intelligent local network that can respond properly to the situation.

The remaining part of this paper is organized as follows. Section II describes the proposed architecture of the Intent-based On-Device AI framework and the main functions of the framework. Section III describes the components and interfaces of the intent-based management framework for On-Device AI. Section IV demonstrates a specific use case to apply this On-Device AI networking framework. Section V concludes this paper along with future work.

II. Components and Interfaces of Intent-based Management Framework for ODAI

This section describes the key components and interfaces of the Intent-based management framework for ODAI. Fig. 1

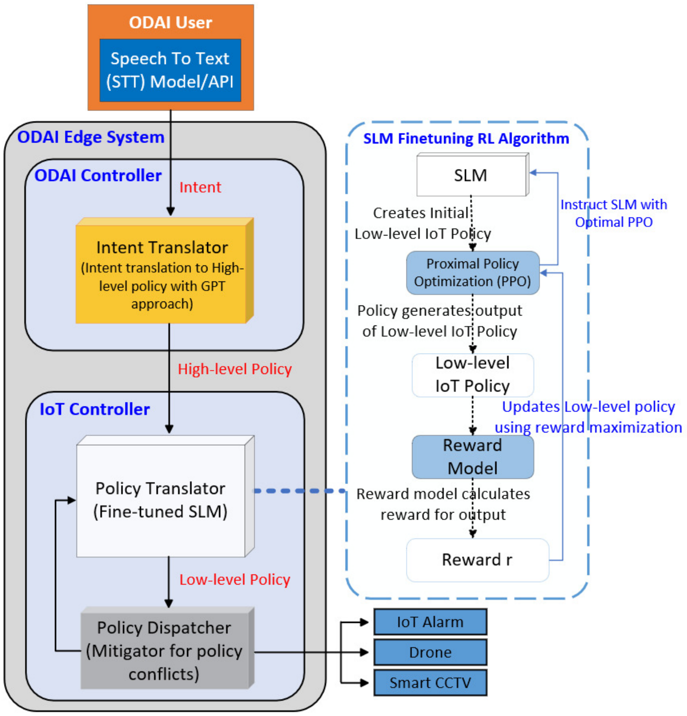
ODAI User
Speech To Text (STT) Model/API
ODAI Edge System
ODAI Controller
Intent
Intent Translator (Intent translation to High-level policy with GPT approach)
High-level Policy
IoT Controller
Policy Translator (Fine-tuned SLM)
Low-level Policy
Policy Dispatcher (Mitigator for policy conflicts)
SLM Finetuning RL Algorithm
SLM
Creates Initial Low-level IoT Policy
Instruct SLM with Optimal PPO
Proximal Policy Optimization (PPO)
Policy generates output of Low-level IoT Policy
Low-level IoT Policy
Updates Low-level policy using reward maximization
Reward Model
Reward model calculates reward for output
Reward r
IoT Alarm
Drone
Smart CCTV

Fig. 2: Architecture of Intent-Based Management Framework for On-Device AI Services

shows the structure of the components and interfaces and indicates how each part is related to other parts. This is a detailed description of the important components of the ODAI management framework.

- **ODAI User:** It is the software (e.g., web-browser-based user interface) used by administrators to deliver network intents to the framework.
- **Knowledge Graph (KG):** It is a knowledge graph database for managing data, including network and security configurations, location data, etc.
- **ODAI Controller (ODAI Translator):** It translates an intent into policies and manages other system components. It interacts with a high-level policy IoT controller via the Controller-Facing Interface.
- **ODAI Vendor's Management System:** It registers service modules and resources with the ODAI Controller via the Registration Interface.
- **ODAI Analyzer:** It gathers monitoring data from IoT devices, evaluates it, and provides feedback for reconfiguration to ODAI Controller via the Analytics Interface.
- **IoT Controller (Policy Translator):** It manages IoT modules, translating high-level policy into a low-level policy that can be executed by the IoT modules (i.e., Service Functions (SF) of the IoT Device).
- **IoT Vendor's Management System:** It registers IoT modules with the IoT Controller via the Registration Interface.
- **IoT Analyzer:** It collects monitoring data from IoT modules, analyzes their performance, and ensures their functionality.

Each component can communicate with the other components through the interfaces. It is necessary to have interfaces between a pair of components on the ODAI management framework. The following is a detailed description of the important interfaces in the ODAI management framework.

- **Consumer-Facing Interface:** Interface between the ODAI User and the ODAI Controller for delivering an intent.
- **Controller-Facing Interface:** Interface between the ODAI Controller and the IoT Controller for transmitting high-level policy.
- **SF-Facing Interface:** Interface between the IoT Controller and the IoT modules (i.e., SFs) for delivering a low-level policy.
- **Registration Interface:** Interface between controllers and vendor management systems for registering and managing components.
- **Monitoring Interface:** Interface between IoT modules and the IoT Analyzer for collecting performance data.
- **Analytics Interface:** Interface for delivering reconfiguration policies and feedback between analyzers and controllers.
- **Analyzer-Facing Interface:** Interface between the ODAI Analyzer and the IoT Analyzer for sharing analysis results.
- **VMS-Facing Interface:** Interface between vendor management systems for exchanging service function information.
- **Database Interface:** Interface for data exchange between the Knowledge Graph (KG) and another component.

III. Architecture of Intent-Based Management Framework for On-Device AI

The architecture of the Intent-Based ODAI Framework is illustrated in Fig. 2, which highlights its key components and interfaces. The framework consists of the ODAI User, Speech-to-Text (STT) Model/API, ODAI Controller (ODAI Translator), IoT Controller (Policy Translator). The primary goal of this component is to enable efficient intent translation and high-policy generation for IoT Controller based on user intent.

The ODAI User provides voice input, which is transmitted to the STT [5] Model on the ODAI User. This model converts the voice input into a textual representation, enabling downstream processing.

The textual message is processed by the ODAI Controller, which employs intent classification to extract and classify the intent from the natural language. Based on the extracted intent, a high-level policy is generated and forwarded to the intent Translator which is a Fine-Tuned SLM (small Language Model) [6].

The Fine-tuned SLM plays a critical role in converting high-level policies into actionable low-level policies tailored for individual IoT devices. The training sequence of the SLM consists of the following steps:

1) **Initial Policy Generation:** The SLM is fine-tuned using Reinforcement Learning (RL) to create an initial low-level IoT policy [7]. Proximal Policy Optimization (PPO) is employed to optimize the policy generation process.
2) **Reward Maximization:** A reward model evaluates the performance of the generated policies by assigning reward values based on their effectiveness. This feedback is used to iteratively improve the low-level policies through RL.
3) **Policy Refinement:** The training loop continues until the low-level policies achieve optimal performance in real-world scenarios. The trained SLM can then generalize across diverse tasks, efficiently converting high-level policies into precise device configurations.

Once the SLM generates a high-level policy, it is sent to the Policy Dispatcher. The Policy Dispatcher ensures that conflicting policies across devices are resolved, and the final, executable high-level policies are delivered to the respective IoT device's (e.g., IoT alarms, drones, and smart CCTVs) controllers.

The Intent-Based ODAI Translator leverages DNN models, efficient training algorithms, and reinforcement learning to enable seamless interaction between users and IoT devices, enhancing their usability and effectiveness in real-world applications.

IV. Use case of Intent-Based Management Framework for On-Device AI

The On-Device AI Framework can be applied to dangerous places (e.g., accident spots) that are difficult for people to reach. Fig. 3 indicates the scene of an indoor emergency. When an emergency occurs, the ODAI User sends an intent to On-Device AI Network (ODAI-Net). Based on the user's intent, the IoT devices collect data and transmit it to the ODAI-Net. ODAI-Net preprocesses and analyzes the data and extracts multimodal features before transmitting the crisis situation to the people inside the room. The intelligent chatbot extracts the intent of the preprocessed data from the multimodal sensors, and transmits the intent to the necessary users. The users can understand the situation and can find an efficient solution in the emergency situation. While the intelligent chatbot informs people of the situation and the evacuation guidance, the On-Device AI SDN (ODAI-SDN) Controller gives instructions to drones and robots, to enable them to react to the situation. The drone or robot scans the specified area for obstacles and notifies the user of them. It also notifies the user of the emergency situation if there is an injured person.

V. Conclusion

In this paper, we introduced the Intent-Based On-Device AI Framework (ODAI) that supports collaborative AI functions between devices based on a user intent. The proposed framework utilizes multimodal sensor data and gives an intent to the IoT devices' components such as an STT Model, Intent Translator, and Policy Coordinator. The framework is explained with a use case focused on emergency response scenarios. The use case shows the capability of the framework to handle critical situations. As future work, our goal is to create more efficient solutions to convey a network intent into an On-Device AI Network, optimizing the management of processes in the network. Additionally, we will build a testbed that can conduct experiments in various IoT environments while integrating advanced machine-learning techniques to verify the scalability and performance of the proposed system.

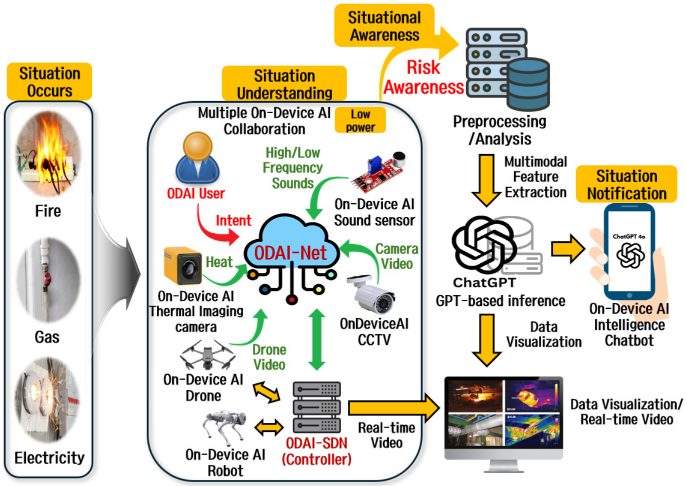

Fig. 3: Use Case of On-Device AI Framework for Safety in Smart Factory

Acknowledgments

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Ministry of Science and ICT (MSIT), South Korea (No. RS-2024-00398199 and RS-2022-II221015). Note that Jaehoon (Paul) Jeong is the corresponding author.

References

[1] R. Gozalo-Brizuela and E. C. Garrido-Merchan, "Chatgpt is not all you need. a state of the art review of large generative ai models," 2023. [Online]. Available: https://arxiv.org/abs/2301.04655

[2] S. Zhu, T. Voigt, F. Rahimian, and J. Ko, "On-device training: A first overview on existing systems," *ACM Transactions on Sensor Networks*, vol. 20, no. 6, p. 1–39, Oct. 2024.

[3] Z. Jia, M. Zaharia, and A. Aiken, "Machine learning systems: Design and implementation," 2023, accessed: 2025-01-05. [Online]. Available: https://mlsysbook.ai/

[4] R. Yang, W. Zhang, N. Tiwari, H. Yan, T. Li, and H. Cheng, "Multimodal sensors with decoupled sensing mechanisms," *Advanced Science*, vol. 9, no. 26, p. 2202470, 2022.

[5] C. Wang, Y. Tang, X. Ma, A. Wu, S. Popuri, D. Okhonko, and J. Pino, "fairseq s2t: Fast speech-to-text modeling with fairseq," 2022. [Online]. Available: https://arxiv.org/abs/2010.05171

[6] L. C. Magister, J. Mallinson, J. Adamek, E. Malmi, and A. Severyn, "Teaching small language models to reason," 2023. [Online]. Available: https://arxiv.org/abs/2212.08410

[7] E. D. S. Pereira, "Opportunities and challenges for tinyml in online distributed learning," May 2023, accessed: 2025-01-05. [Online]. Available: https://cms.tinyml.org/wp-content/uploads/talks2023/tinyML_ODL_Forum_Eduardo_Dos_Santos_Pereira_230516.pdf