RIDAS: A Multi-Agent Framework for AI-RAN with Representation- and Intention-Driven Agents

Kuiyuan Ding[*], Caili Guo[†], Yang Yang[*] and Jianzhang Guo[‡]
[*] Beijing Key Laboratory of Network System Architecture and Convergence, School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China
[†] Beijing Laboratory of Advanced Information Networks, School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China
[‡] China Telecom Digital Technology Co., Ltd., Beijing 100035, China
{dingkuiyuan, guocaili, yangyang01}@bupt.edu.cn, guojz6@chinatelecom.cn

***Abstract*—Sixth-generation (6G) networks demand tight integration of artificial intelligence (AI) into radio access networks (RANs) to meet stringent quality-of-service (QoS) and resource-efficiency requirements. Existing solutions struggle to bridge the gap between high-level user intents and the low-level, parameterized configurations required for optimal performance. To address this challenge, we propose RIDAS, a multi-agent framework composed of representation-driven agents (RDAs) and an intention-driven agent (IDA). RDAs expose open interface with tunable control parameters—rank and quantization bits, enabling explicit trade-offs between distortion and transmission rate. The IDA employs a two-stage planning scheme (bandwidth pre-allocation and reallocation) driven by a large language model (LLM) to map user intents and system state into optimal RDA configurations. Experiments demonstrate that RIDAS supports 44.71% more users than WirelessAgent under equivalent QoS constraints. These results validate ability of RIDAS to capture user intent and allocate resources more efficiently in AI-RAN environments. Code is available on: https://github.com/echojayne/RIDAS.git**

***Index Terms*—6G, AI-RAN, AI agents, resource allocation, large language models.**

I. Introduction

Sixth-generation (6G) networks envision a profound integration of artificial intelligence (AI) into communication infrastructures, thereby imposing more stringent requirements on the radio access network (RAN). In response, the concept of AI-enabled RAN (AI-RAN) has emerged, leveraging advanced AI methodologies to endow the RAN with enhanced intelligence—enabling higher resource utilization and improved quality of service (QoS) [1]. AI-RAN aspires to automate network management by translating high-level business objectives or user intents into concrete network configurations and policy directives. However, the use of conventional AI techniques in AI-RAN reveals a substantial gap between user intents, typically expressed in natural language, and the complex, parameterized configurations required for RAN deployment.

As AI technologies and hardware have advanced, LLMs have demonstrated exceptional performance in general-purpose domains. Their strong capability for intent understanding makes them promising candidates for enhancing AI-RAN. However, standalone LLMs face three key limitations: They cannot efficiently process multimodal data, dynamically decompose complex tasks, or interface with specialized tools, all of which impede their deployment in complex wireless environments [2].

To overcome these obstacles, LLM-based agents have emerged. By augmenting LLMs with modular functionalities—such as environmental data perception and external tool integration—these agents can interpret complex wireless contexts, make informed decisions, and execute appropriate actions. There has been many preliminary explorations that engage LLM-based agents in RAN. The authors in [3] proposed the LLM-empowered hierarchical RAN intelligent controllers (RICs) (LLM-hRIC) framework to improve the collaboration between RICs in open RAN. In [2], the authors introduce a framework called WirelessAgent harnessing LLMs to create autonomous AI agents for diverse wireless network tasks.

Despite LLM-based agents advances, existing studies largely overlook how high-level, intention-driven LLM agents which serves as the “brain” of AI-RAN can effectively orchestrate underlying AI models, particularly those for data representation that lack natural-language capabilities. Specifically, there are two primary challenges that must be addressed:

- Challenge 1: How can underlying AI models be designed with an open interface that exposes tunable operational parameters, enabling external, high-level control over the trade-off between resource efficiency and QoS?
- Challenge 2: Given such a controllable interface, what framework should an LLM-based agent employ to translate high-level user intents into a sequence of concrete control actions, dynamically adapting to network status to orchestrate the underlying models?

In this article, we propose RIDAS, a multi-agent framework for AI-RAN that consists of representation-driven agents (RDAs) and an intention-driven agent (IDA). RIDAS enables the RAN to allocate bandwidth resources efficiently so as to serve as many users as possible while satisfying their QoS (task performance) requirements. Specifically, we first design the RDA, deployed at the user end, which employs sign-

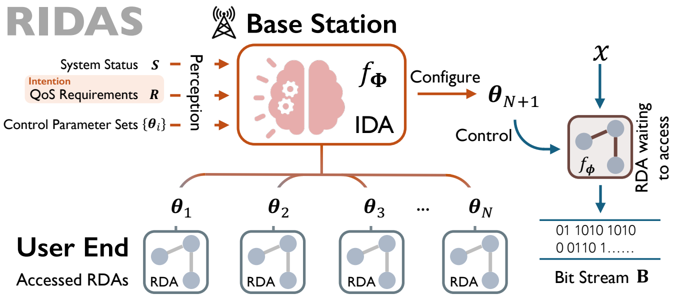


Fig. 1: Overall architecture of proposed RIDAS framework.

value-independent decomposition (SVID) [4] to represent the source; by adjusting the decomposition rank and the number of quantized bits, both the transmission rate and the QoS performance of the RDA can be precisely controlled. Second, we introduce the IDA, which takes the QoS requirements of users as the intention input and employs a two-stage planning scheme—bandwidth pre-allocation and bandwidth reallocation—to dynamically adjust the control parameters of RDAs, thereby minimizing bandwidth consumption while meeting user QoS demands. Experimental results show that RIDAS can support 44.71% more users than the WirelessAgent framework under equivalent QoS constraints, demonstrating that RIDAS effectively captures user intent and allocates system resources more efficiently.

II. System Model and Problem Formulation

The overall architecture of RIDAS framework is shown in Fig. 1. At the user end (UE), there are N RDAs connected to the base station (BS), each governed by a distinct control parameter $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_N$. When a new RDA attempts to access the BS, the IDA, deployed at the BS, determines its control parameter $\boldsymbol{\theta}_{N+1}$ based on the current system state S, the parameter sets of the existing RDAs $\{\boldsymbol{\theta}_i \mid i = 1, \cdots, N\}$ and the specific requirements R of the new RDA (e.g., QoS constraints). Denoting the IDA as a mapping function $f_{\boldsymbol{\Phi}}$ parameterized by $\boldsymbol{\Phi}$, this process can be formulated as follows:

$$\boldsymbol{\theta}_{N+1} = f_{\boldsymbol{\Phi}}\big(S, R, \{\boldsymbol{\theta}_i\}\big). \tag{1}$$

The configured control parameter $\boldsymbol{\theta}_{N+1}$ governs how the new RDA represents messages. Specifically, when a message x needs to be encoded, the RDA denoted by $f_{\boldsymbol{\phi}}$ and parameterized by $\boldsymbol{\phi}$ generates the encoded bit stream $\mathcal{B}$ under the guidance of $\boldsymbol{\theta}_{N+1}$. This process is formulated as:

$$\mathcal{B} = f_{\boldsymbol{\phi}}(x; \boldsymbol{\theta}), \tag{2}$$

where $\mathcal{B}$ represents the output bit stream.

The design of the RDA should ensure that both the quality and the quantity of the generated representations can be effectively controlled by the associated control parameter $\boldsymbol{\theta}_{N+1}$. Here, the quality of the representations reflects the level of distortion, while the quantity measured by the length of the encoded bit stream $\mathcal{B}$ corresponds to the transmission rate. Therefore, Challenge 1 lies in identifying a feasible mapping that jointly satisfies these objectives by balancing representation fidelity and communication efficiency.

The objective of designing the IDA is to determine an optimal control parameter $\boldsymbol{\theta}_{N+1}$ that achieves a balance between transmission rate and distortion. This goal can be formulated as the following constrained optimization problem:

$$\begin{aligned} \boldsymbol{\theta}^*_{N+1} = {} & \arg\min_{f_{\boldsymbol{\Phi}}} \underbrace{\mathbb{E}_{x\sim p_X}\big[\,|\mathcal{B}|\,\big]}_{\mathcal{R}(\boldsymbol{\theta}_{N+1})} \\ \text{s.t.} \quad & \sum_{i=1}^{N+1} \mathcal{R}(\boldsymbol{\theta}_i) \;\le\; B_{\max}(S), \\ & \mathcal{B} = f_{\boldsymbol{\phi}}(x; \boldsymbol{\theta}_{N+1}), \\ & D(\boldsymbol{\theta}_{N+1}) \;\le\; D_{\text{req}}(R), \\ & \boldsymbol{\theta}_{N+1} = f_{\boldsymbol{\Phi}}\big(S, R, \{\boldsymbol{\theta}_i, i = 1, \cdots, N\}\big), \end{aligned} \tag{3}$$

where $D(\boldsymbol{\theta})$ and $\mathcal{R}(\boldsymbol{\theta})$ denote the average distortion and rate under the control parameter $\boldsymbol{\theta}$, respectively, $B_{\max}(S)$ is the total bandwidth budget imposed by the current system state S, $D_{\text{req}}(R)$ is the distortion requirement determined by the QoS constraint R of the new RDA and $|\mathcal{B}|$ denotes the length of the generated bit stream $\mathcal{B}$.

By solving the optimization problem in Eq. (3) through the design of a suitable IDA function $f_{\boldsymbol{\Phi}}$, we aim to effectively address Challenge 2.

III. Proposed RIDAS Framework

The overall architecture of the proposed RIDAS framework is illustrated in Fig. 1 which comprises two primary components: RDAs at the user end and an IDA at BS. The RDAs expose interfaces parameterized by $\boldsymbol{\theta}$, which the IDA configures by computing a near-optimal control parameter $\boldsymbol{\theta}^*$. This process ensures that user QoS requirements are satisfied while minimizing system resource utilization.

In this section, we first present the design of the RDAs, and then provide a detailed description of the overall IDA architecture.

A. The design of RDA

As illustrated in Fig. 2, the RDA architecture comprises a well-trained deep neural network (DNN) denoted as $g_{\boldsymbol{\phi}}$ parameterized by $\boldsymbol{\phi}$, followed by the SVID module and subsequent quantization and entropy-coding stages.

1) DNN: The DNN within the RDA serves as the core component for data representation and fundamentally determines its effectiveness. In our framework, we abstract away the specifics of the DNN architecture and training procedure, assuming that it produces sufficiently efficient representations. Instead, we concentrate on the downstream processing of these representations, providing a programmable interface that enables precise control over both representation quality and transmission efficiency.

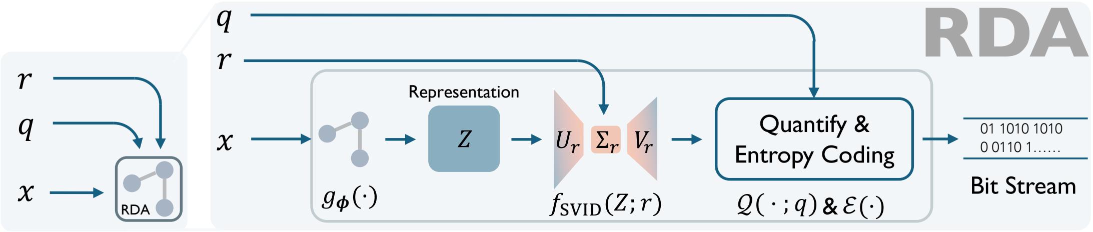


Fig. 2: The architecture of RDA.

2) SVID: The representation space produced by the DNN is typically very high-dimensional, and directly transmitting these raw representations would incur a substantial bandwidth overhead. To mitigate this, we innovatively leverage the SVID method introduced in [4], which was originally developed for decomposing DNN weights. Concretely, let the output representation of the DNN be denoted as $Z \in \mathbb{R}^{m\times n}$, and choose a target rank $r \ll \min(m,n)$. We then form the following rank-r approximation:

$$Z \approx Z_{\text{sign}} \odot \left(U_r \Sigma_r V_r^T\right), \tag{4}$$

In Eq. (4), $Z_{\text{sign}} = \text{sign}(Z) \in \{-1,1\}^{m\times n}$ is the element-wise sign matrix of Z, where

$$\text{sign}(z_{ij}) = \begin{cases} +1, & z_{ij} \geq 0, \\ -1, & z_{ij} < 0. \end{cases}$$

The operator $\odot$ denotes the Hadamard (elementwise) product. The factorization $U_r \Sigma_r V_r^T$ is the first r singular value decomposition (SVD) of $|Z|$, the matrix obtained by taking the absolute value of each entry of Z. The matrices $U_r \in \mathbb{R}^{m\times r}$ and $V_r \in \mathbb{R}^{r\times n}$ are the matrices of the first r left and right singular vectors of $|Z|$, respectively. Finally, $\Sigma_r = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r) \in \mathbb{R}^{r\times r}$ is the diagonal matrix of singular values ordered as $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq 0$.

The following proposition is the reason why we choose SVID over the conventional SVD method.

Proposition 1. *Given a matrix* $\mathbf{W}$ *and its element-wise absolute value* $|\mathbf{W}|$*, let*

$$\mathbf{W} = \mathbf{W}_{\text{sign}} \odot |\mathbf{W}|.$$

We decompose these as

$$\mathbf{W} = \mathbf{a}\mathbf{b}^\top + \mathbf{E}_1, \quad |\mathbf{W}| = \tilde{\mathbf{a}}\tilde{\mathbf{b}}^\top + \mathbf{E}_2,$$

where a, b *and* $\tilde{a}, \tilde{b}$ *are vectors of appropriate dimensions given by SVD, each* $\mathbf{E}_i$ *is the corresponding error matrix. In terms of the Frobenius norm, the SVID approximation is closer to the original matrix* $\mathbf{W}$*:*

$$\left\|\mathbf{W} - \mathbf{W}_{\text{sign}} \odot \tilde{\mathbf{a}}\tilde{\mathbf{b}}^\top\right\|_F^2 \leq \left\|\mathbf{W} - \mathbf{a}\mathbf{b}^\top\right\|_F^2. \tag{5}$$

Proof. For a detailed proof of this proposition, we refer the reader to the work of Xu et al. [4]. □

As shown in Eq.(5), the reconstruction error of SVID is provably no greater than that of SVD. Furthermore, Proposition 1 can be straightforwardly generalized to rank-r decompositions. Accordingly, SVID not only achieves a lower-error approximation $U_r \Sigma_r V_r^T$ than SVD, but also offers a controllable trade-off between representation dimension and distortion, thereby providing a preliminary solution to Challenge 1.

3) Quantization and Entropy-Coding: For digital transmission, we further apply quantization and entropy encoding to the matrices U_r, Σ_r, and V_r^T. Specifically, we denote the SVID-based low-rank approximation as $f_{\text{SVID}}(Z;r)$, the quantization process as $\mathcal{Q}(W;q)$ where q indicates the number of quantization bits, and the subsequent entropy encoding as $\mathcal{E}(W_{\text{Q}})$, where W_{Q} denotes the quantized representation. Accordingly, the RDA parameterized by ϕ can be expressed as:

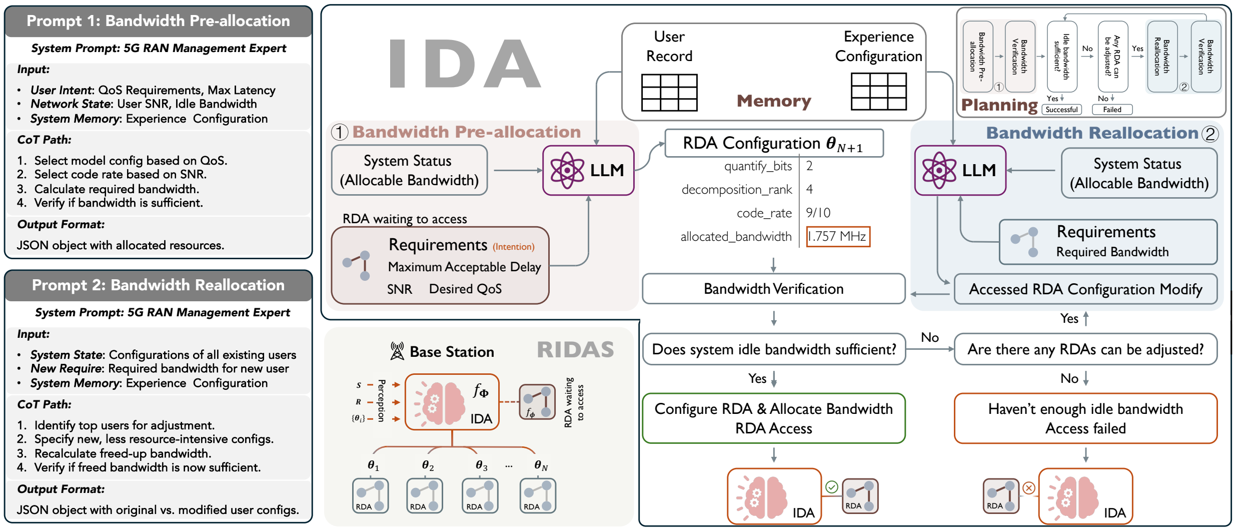
$$\begin{aligned} \mathcal{B} &= f_{\phi}(x;\boldsymbol{\theta}) \\ &= \mathcal{E}\big(\mathcal{Q}\big(f_{\text{SVID}}\big(g_{\phi}(x);r\big);q\big)\big), \end{aligned} \tag{6}$$

where the control parameter $\boldsymbol{\theta} = \{r, q\}$.

Eq. (6) presents a feasible implementation of the RDA. The distortion and transmission rate of the resulting representation bit stream can be effectively controlled by rank r and quantization bits q. Specifically, larger values of r and q lead to higher transmission rates but lower task performance distortion, while smaller values of r and q result in reduced transmission rates at the cost of increased task performance distortion. Thus, this implementation provides a controllable trade-off between resource efficiency and QoS. Up to this point, we have presented a viable instantiation of the RDA, thereby addressing Challenge 1.

B. The design of IDA

As illustrated in Eq. (3), IDA seeks to determine the optimal control parameter $\boldsymbol{\theta}_{N+1}^*$ for the new RDA, such that the resulting representation bit stream length $|\mathcal{B}|$ is minimized while satisfying the maximum-distortion constraint of the RDA. As illustrate in Fig. 3, we realize IDA as an LLM-based agent composed of two principal routines: (1) bandwidth pre-allocation and (2) bandwidth reallocation. Specifically, the IDA first selects $\boldsymbol{\theta}_{N+1}^*$ for the new RDA and then allocates the corresponding transmission bandwidth. In the following subsections, we first present the memory module of IDA, and then describe its planning pathway in detail.



Prompt 1: Bandwidth Pre-allocation

System Prompt: 5G RAN Management Expert

Input:

- *User Intent*: QoS Requirements, Max Latency
- *Network State*: User SNR, Idle Bandwidth
- *System Memory*: Experience Configuration

CoT Path:

1. Select model config based on QoS.
2. Select code rate based on SNR.
3. Calculate required bandwidth.
4. Verify if bandwidth is sufficient.

Output Format:

JSON object with allocated resources.

Prompt 2: Bandwidth Reallocation

System Prompt: 5G RAN Management Expert

Input:

- *System State*: Configurations of all existing users
- *New Require*: Required bandwidth for new user
- *System Memory*: Experience Configuration

CoT Path:

1. Identify top users for adjustment.
2. Specify new, less resource-intensive configs.
3. Recalculate freed-up bandwidth.
4. Verify if freed bandwidth is now sufficient.

Output Format:

JSON object with original vs. modified user configs.

Fig. 3: Overall architecture of proposed IDA.

1) Memory: The memory module of IDA persistently stores the current configuration set, i.e., the control parameters of all connected RDAs $\boldsymbol{\theta}_i$, together with the allocated bandwidth of each RDA (*User Record* table in Fig. 3) and the corresponding achieved experimental bit stream length and distortion under those configurations (*Experience Configuration* table in Fig. 3). All of this information is maintained in a real-time updating database, which serves as the perception input to the LLM-based agent.

2) Planning: In order to enable the LLM to generate near-optimal configurations, we have designed a dedicated planning pathway for IDA.

When a new RDA requests access to the BS, the LLM within IDA first proposes control parameter $\boldsymbol{\theta}_{N+1}$ and pre-allocates bandwidth according to the QoS requirements of the RDA (including distortion and rate requirements) and the currently available idle bandwidth of the system. During this pre-allocation stage, the prompt steers the LLM to retrieve from past experience a configuration that minimizes transmission rate while satisfying the distortion constraint. Owing to its strong instruction-following and contextual-understanding capabilities, the initial proposal of the LLM typically lies very close to the true optimum, i.e., $\boldsymbol{\theta}_{N+1} \approx \boldsymbol{\theta}^*_{N+1}$, and thus the resulting bandwidth allocation effectively balances system resource usage against user demand.

However, since LLM are prone to numerical hallucinations when performing precise calculations, we further validate the pre-allocated bandwidth by measuring the empirical transmission rate under the proposed configuration before committing to the final assignment.

If the idle bandwidth of system is insufficient to satisfy the pre-allocation request, IDA initiates a reallocation procedure across all RDAs connected to the BS rather than rejecting the new RDA directly. Because the control parameters and bandwidth assignments determined during the pre-allocation stage may not be optimal, some RDAs may hold redundant capacity. Consequently, when idle bandwidth is inadequate, the LLM of IDA evaluates whether the configuration of any connected RDA can be adjusted to free up additional resources. If such opportunities exist, IDA modifies those configurations and reallocates their bandwidth. It then reassesses the available bandwidth. If the newly available capacity meets the requirements, the new RDA is configured, and the corresponding bandwidth is provisioned for it. Otherwise, the reallocation stage repeats until either sufficient idle bandwidth is secured (connected successfully) or no further adjustments are possible (connected failed).

By employing the two-stage planning pathway of bandwidth pre-allocation and reallocation, IDA is able to assign near-optimal control parameters and bandwidth resources so as to minimize the representation bit-stream length $|\mathcal{B}|$. This not only conserves system resources but also ensures that the QoS requirements of each RDA are met, thereby directly addressing the optimization objective in Eq. (3). The concrete instantiation of IDA presented here thus constitutes a viable solution to Challenge 2.

The proposed RDA and IDA together form the RIDAS framework. On the one hand, RDAs expose an interface for controlling their representation actions, thereby satisfying diverse user QoS requirements with minimal resource consumption through the adjustment of the control parameter $\boldsymbol{\theta}$. On the other hand, the IDA interprets user intent to configure the control parameters of RDAs in a manner that further minimizes system resource usage. Consequently, RIDAS introduces a novel AI-RAN paradigm, demonstrating how LLM-based agents can interact with and manage the underlying AI models.

IV. Experimental Results

A. Settings

1) Scenario setup: In our experiments, the total available bandwidth is set to 100 MHz. The signal-to-noise ratio (SNR) for each RDA is randomly generated in the range of 5 dB to 30 dB. The RDA code rate is selected from the set

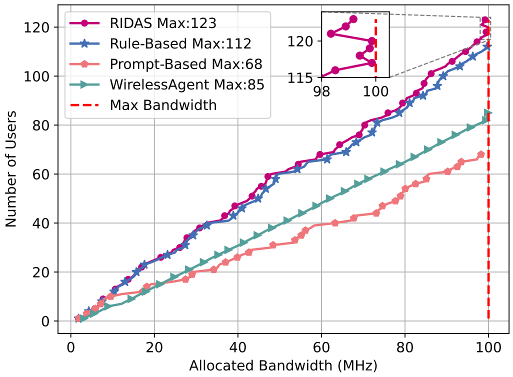


Fig. 4: Number of users connected to BS of different methods with the same bandwidth.

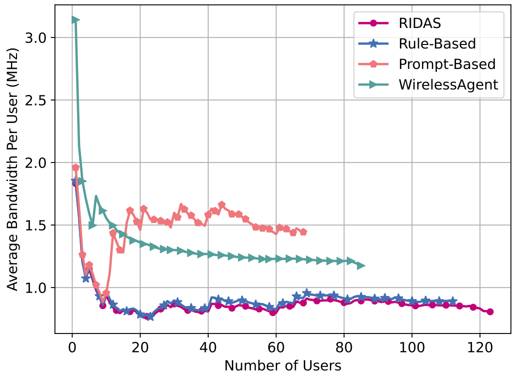


Fig. 5: Average bandwidth of connected users.

$\{\frac{1}{2}, \frac{3}{5}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}, \frac{5}{6}, \frac{8}{9}, \frac{9}{10}\}$. According to [5], the end-to-end delay in 6G networks is expected to be less than 1 ms; accordingly, we impose a maximum allowable transmission delay per RDA in the interval [0.05 ms, 0.5 ms]. Moreover, we characterize QoS requirement of each RDA by its top-1 classification accuracy:

- Low: QoS demands accuracy $> 70\%$, corresponding to shorter bit-stream lengths;
- Medium: QoS requires accuracy $> 80\%$;
- High: QoS requires accuracy $> 90\%$, corresponding to longer bit-stream lengths.

Under these configurations, the required bandwidth is computed as follows:

$$\text{Bandwidth (MHz)} = \frac{|\mathcal{B}| \,/\, \text{code rate}}{\text{transmission time}} \times \frac{1}{\log_2\left(1 + 10^{\frac{\text{SNR}}{10}}\right) \times 10^6}. \tag{7}$$

2) Task and model setup: To evaluate the effectiveness of our RDA design, we perform an image-classification task on the CIFAR-10 dataset [6], using a ViT-B/16 architecture from the CLIP framework [7] as the representation backbone g_ϕ. The LLM served in IDA is set to DeepSeek-V3-0324 [8].

B. Baselines

For fair comparisons, we employ the following baselines:

- WirelessAgent [2]: An LLM-based autonomous agent for wireless tasks.
- Prompt-Based: A simplified version of our method that uses a single LLM prompt for allocation, omitting the verification and reallocation stages.
- Rule-Based: A heuristic approach that allocates bandwidth based on optimal control parameters and an SNR-scaled code rate.

C. Results

We simulate a common queue of users awaiting connection across all methods. Hence, under a fixed total bandwidth, supporting a greater number of users or, for the same number of connected users, allocating a lower average bandwidth to each user while still meeting their QoS requirements indicates more efficient resource utilization.

Fig. 4 shows the number of users that can be supported by the BS under different baseline methods, given an identical total bandwidth budget. As illustrated in Fig. 4, when the allocated bandwidth approaches 100% of the system capacity, the proposed RIDAS accommodates up to 123 users, compared with 112 for the rule-based baseline, 85 for the WirelessAgent framework, and 68 for the prompt-based scheme.

These results indicate that the proposed IDA not only attains near-optimal control-parameter configurations but also supports a greater number of concurrent users than the rule-based scheme. This improvement stems from the fact that the rule-based method selects code rates by linearly scaling with SNR, whereas IDA adaptively determines code rates based on empirical performance data—thereby conserving bandwidth more effectively across diverse scenarios.

Furthermore, for RIDAS, when 117 users are connected, the allocated bandwidth is nearly exhausted. Upon the 118th connection attempt, IDA triggers its reallocation stage to reclaim additional bandwidth from existing RDAs, thereby freeing sufficient capacity to admit the additional user.

Fig. 5 shows the average bandwidth allocated per user as the connected users varies. At 85 concurrent users, the mean per-user allocation is 0.893 MHz under RIDAS, compared with 0.925 MHz for Rule-Based method and and 1.175 MHz for WirelessAgent framework. These results further demonstrate that the proposed IDA distributes bandwidth more efficiently while still meeting QoS requirements of each user.

Fig. 6 details the per-user bandwidth assignments for varying acceptable transmission-latency requirements across different number of connected user. As shown in Fig. 6 , RIDAS dynamically adapts allocation according of each RDA to its latency tolerance—users that can tolerate higher delays are provisioned with less bandwidth, whereas those requiring lower latency receive larger allocations—thereby conserving resources. By contrast, the WirelessAgent framework tends to distribute bandwidth uniformly among all users. These findings further demonstrate the ability of IDA within RIDAS to infer the intent of each agent and generate control-parameter configurations that optimize resource utilization.

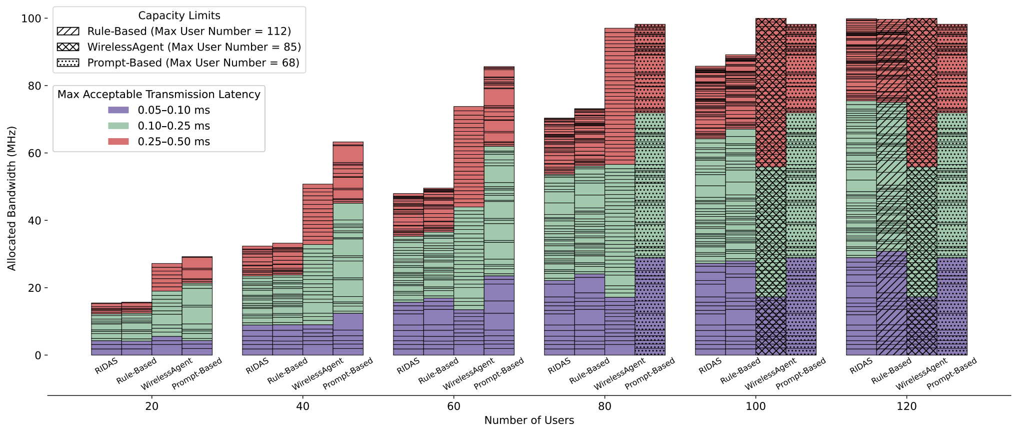


Fig. 6: Details of bandwidth allocation.

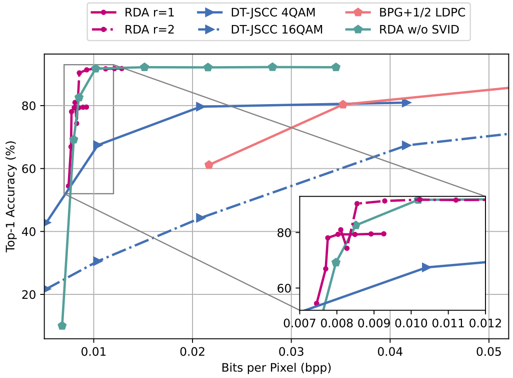


Fig. 7: Accuracy versus bpp.

To demonstrate the effectiveness of the proposed RDA, we compare its performance against two key baselines. Fig. 7 illustrates this comparison by plotting the classification accuracy as a function of bits per pixel (bpp)[1]. The baselines used for this comparison are DT-JSCC [9], a task-oriented communication scheme employing digital modulation, and BPG + 1/2 LDPC, which represents a conventional approach of transmitting images using BPG coding protected by a rate-1/2 LDPC code. As shown in Fig. 7, at approximately 0.008 bpp, RDA outperforms DT-JSCC with 4-QAM and the RDA variant without SVID by over 20% and 18% in accuracy, respectively. These results demonstrate that RDA effectively preserves the quality of the representation—and hence the downstream task performance—even in the low-bpp regime. At higher bpp values, where representational distortion becomes negligible, accuracy of RDA approaches that of the original representation model.

V. Conclusion

In this work, we have introduced RIDAS, a multi-agent framework for AI-RAN that unifies low-level representation control with high-level intent interpretation via its RDA and IDA components. In our evaluation, RIDAS dynamically adjusted its control parameters in response to network conditions and user QoS requirements, thereby maximizing resource utilization. Its two-stage planning process, comprising bandwidth pre-allocation and subsequent reallocation, achieved near-optimal performance by satisfying both transmission-rate and task-performance demands. As a contribution, RIDAS offers a novel paradigm for autonomous, intent-driven RAN management and provides a promising foundation for 6G networks. Future work will extend the framework to incorporate end-to-end delay control and adapt to more complex deployment scenarios, further enhancing its practicality in real-world wireless environments.

References

[1] L. Kundu, X. Lin, R. Gadiyar, J.-F. Lacasse, and S. Chowdhury, “Ai-ran: Transforming ran with ai-driven computing infrastructure,” *https://arxiv.org/abs/2501.09007*, January. 2025.

[2] J. Tong, J. Shao, Q. Wu, W. Guo, Z. Li, Z. Lin, and J. Zhang, “Wirelessagent: Large language model agents for intelligent wireless networks,” *https://arxiv.org/abs/2409.07964*, July. 2024.

[3] L. Bao, S. Yun, J. Lee, and T. Q. S. Quek, “Llm-hric: Llm-empowered hierarchical ran intelligent control for o-ran,” *https://arxiv.org/abs/2504.18062*, April. 2025.

[4] Y. Xu, X. Han, Z. Yang, S. Wang, Q. Zhu, Z. Liu, W. Liu, and W. Che, “Onebit: Towards extremely low-bit large language models,” *https://arxiv.org/abs/2402.11295*, February. 2024.

[5] A. Salh, L. Audah, N. S. M. Shah, A. Alhammadi, Q. Abdullah, Y. H. Kim, S. A. Al-Gailani, S. A. Hamzah, B. A. F. Esmail, and A. A. Almohammedi, “A survey on deep learning for ultra-reliable and low-latency communications challenges on 6g wireless systems,” *IEEE Access*, vol. 9, pp. 55 098–55 131, March. 2021.

[6] A. Krizhevsky, G. Hinton *et al.*, “Learning multiple layers of features from tiny images,” *University of Toronto Tech. Rep*, vol. 1, January. 2009.

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning transferable visual models from natural language supervision,” *arXiv preprint arXiv:2103.00020*, March. 2021.

[8] B. F. DeepSeek-AI, Aixin Liu and et al., “Deepseek-v3 technical report,” *https://arxiv.org/abs/2412.19437*, December. 2025.

[9] S. Xie, S. Ma, M. Ding, Y. Shi, M. Tang, and Y. Wu, “Robust information bottleneck for task-oriented communication with digital modulation,” *IEEE J. Sel. Areas Commun.*, vol. 41, no. 8, pp. 2577–2591, June. 2023.

[1] Here, bpp $= \frac{\text{total transmission bits}}{3\times\text{Height}\times\text{Width}}$, where Height and Width denote the image dimensions.