Low-latency MLLM Inference with Spatiotemporal Heterogeneous Distributed Multimodal Data

Xiangrui Xu[a], Sicong Liu[a,*], Zhiwen Yu[a,b,*], Lehao Wang[a], Bin Guo[a]
[a] School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[b] Harbin Engineering University, Harbin, China
{xiangruixu}@mail.nwpu.edu.cn, {scliu, zhiwenyu}@nwpu.edu.cn
{lehaowang}@mail.nwpu.edu.cn, {guob}@nwpu.edu.cn

***Abstract*—Distributed sensing systems have been widely applied in various Internet of Things (IoT) scenarios, and the emergence of the Multimodal Large Language Model (MLLM) has opened up new possibilities for these systems. However, the spatiotemporal heterogeneity and asynchronous arrival of distributed mobile data make achieving low-latency, high-accuracy MLLM inference extremely challenging. In this paper, we propose a framework of MLLM inference with spatiotemporal heterogeneous distributed data to achieve low-latency, high-accuracy MLLM inference in distributed sensing systems.**

***Index Terms*—distributed multimodal system, MLLM, multimodal inference,low-latency inference**

I. Introduction

With the proliferation of low-power yet data-rich edge sensing devices, distributed multi-sensor perception technology has been widely applied in various scenarios, such as human activity recognition, autonomous driving, and event detection. Nevertheless, in dynamic and complex distributed scenarios, multimodal models encounter challenges like data drift, multimodal data fusion difficulties, and poor scalability, hindering meeting the essential requirements of the system.

The advent of MLLM has introduced novel opportunities for distributed perception systems. A potential approach entails harnessing LLMs as foundational knowledge and a cognitive driving force to augment the performance of multimodal inference. Compared with traditional multimodal models, MLLM has the following advantages in distributed sensing systems: **(i) Cross-modal fusion understanding.** MLLM provides a unified fusion framework for various types of mobile perception data and leverages prior knowledge and the cognitive driving force of LLM to achieve efficient cross-modal fusion learning. **(ii) Context awareness and inferential abilities.** MLLM can integrate contextual information obtained from mobile devices with real-time sensing data for accurate and efficient inference. **(iii) Generalization and extensible architecture.** Adapting MLLM to various multimodal inference tasks only requires adding or removing modal encoders and fine-tuning the model.

However, in practical distributed multimodal perception systems, using MLLM for low-latency, high-accuracy multimodal inference is extremely challenging due to its sensitivity to sensor data quality. On one hand, due to the distributed deployment and instability of sensors, it can result in spatiotemporal heterogeneity of sensor data and information loss or corruption. On the other hand, due to variations in perception rates, differences in data scales, and random network fluctuations, there are issues of data asynchrony and loss during the distributed data to edge servers aggregation process. The spatiotemporal heterogeneity and information loss or corruption in sensor data increases the difficulty of modality alignment, potentially resulting in hallucination and a significant decrease in accuracy [1], and data arriving asynchronously will increase the waiting delay. Furthermore, there are two major challeges in existing non-blocking inference techniques:

- Most distributed sensor spatiotemporal alignment methods are designed for specific modal combinations. They are deeply coupled with multimodal models, making it difficult to generalize and apply to other modal combinations or DNN models. Also, most generic alignment methods focus on temporal alignment, without considering the spatial heterogeneity. Due to the complex and modular structure and diverse modal combinations in MLLM, developing resource-efficient and generic distributed sensor data spatiotemporal alignment methods is exceedingly challenging.
- Under conditions of asynchronous data arrival, performing inference without waiting for complete data can reduce accuracy, while waiting for all data increases system latency. Existing data imputation methods using generative models and matrix factorization need help to handle changing sensor data in real-time [2], [3]. This limitation may result in hallucinations and decreased accuracy in inference for MLLM. Exploring more applicable missing data imputation methods to overcome the problems of accuracy degradation in MLLM is another major challenge.

To address these challenges, this paper introduces a low-latency and high-accuracy MLLM inference framework in distributed sensing systems. The novelty of this framework lies in proposing mobile sensing data quality enhancement and intermediate result rollback computation schemes to overcome the spatiotemporal heterogeneity and asynchronous arrival of data.

*Corresponding author: scliu@nwpu.edu.cn, zhiwenyu@nwpu.edu.cn

979-8-3503-6338-8/24/$31.00 ©2024 IEEE
DOI 10.1109/CSCAIoT62585.2024.00009

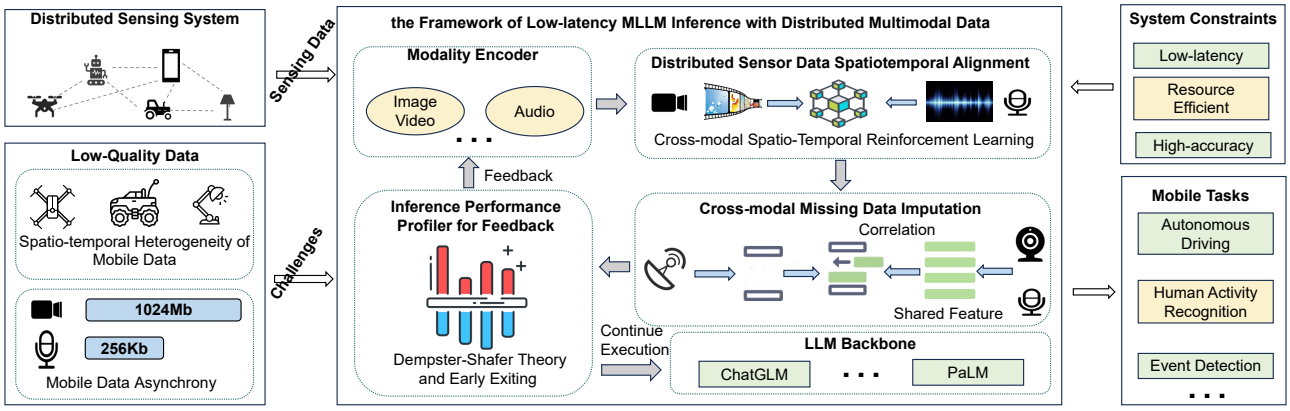

Fig. 1. the framework of low-latency MLLM inference with spatiotemporal heterogeneous distributed multimodal data.

II. System Design

The MLLM inference framework mainly includes the distributed data spatiotemporal alignment module, the cross-modal missing data imputation module, and the inference performance profiler for feedback module. The following will provide a detailed explanation of the components of the MLLM inference framework.

The **distributed sensor data spatiotemporal alignment** module utilizes the spatiotemporal consistency between sensor modalities to address the spatiotemporal heterogeneity of data. As temporal asynchrony exacerbates the spatial heterogeneity of distributed sensing data, we employ a Transformer-based approach to learn the temporal information between modalities and compute the Dynamic Time Warping distance to find the optimal delay configuration. Then, we propose a spatiotemporal enhancement learning method to enhance the spatiotemporal correlation between different modalities, thereby alleviating spatiotemporal heterogeneity and facilitating subsequent precise semantic alignment.

The **cross-modal missing data imputation** module is proposed to address the high latency in MLLM inference caused by asynchronous data arrival. Based on the complementarity between different sensor data, interpolate asynchronously delayed data to achieve high-accuracy MLLM inference without waiting for new data to arrive at the edge server. We employ a shared feature-based approach to capture cross-modal consistency information for each modality. Additionally, we innovatively introduce parameters for historical modality correlations, real-time modality correlations, and modality dynamic changes to perceive the variability of inter-modality correlations and select the optimal shared feature strategy to the imputation of missing modal data.

The **inference performance profiler for feedback** module is proposed to ensure the accuracy of MLLM inference and further reduce the latency. The key idea is to conduct a lightweight analysis of the aligned features to determine whether the current inference should be rolled back and wait for all data to arrive, continue with MLLM inference, or directly output intermediate inference results. Drawing inspiration from early-exit strategies in models, after performing multimodal alignment in MLLM, we can conduct intermediate result inference using available multimodal features. Furthermore, by leveraging the Dempster-Shafer theory to acquire prior information, we partition it into three belief spaces, trust, doubt, and rejection, to evaluate intermediate inference results to achieve low-latency and high-accuracy MLLM inference.

III. Conclusion

This paper presents a framework for low-latency MLLM inference using spatiotemporal heterogeneous distributed multimodal data.The system is divided into three main modules: the distributed sensor data spatiotemporal alignment module, the cross-modal missing data imputation module, and the inference performance profiler for feedback module. These modules optimize the distributed multimodal system to overcome the heterogeneity and asynchronous arrival of distributed sensor data to achieve high-accuracy, low-latency MLLM inference.

Acknowledgment

This work was supported in part by the National Science Fund for Distinguished Young Scholars (No. 62025205), the National Key RD Program of China (No. 2021YFB2900100), and the National Natural Science Foundation of China (No. 61960206008, No. 62032020, No. 62102317).

References

[1] S. Song, X. Li, and S. Li, "How to bridge the gap between modalities: A comprehensive survey on multimodal large language model," *arXiv preprint arXiv:2311.07594*, 2023.

[2] J. Wang, G. Wang, X. Zhang, L. Liu, H. Zeng, L. Xiao, Z. Cao, L. Gu, and T. Li, "Patch: A plug-in framework of non-blocking inference for distributed multimodal system," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 7, no. 3, pp. 1–24, 2023.

[3] T. Li, J. Huang, E. Risinger, and D. Ganesan, "Low-latency speculative inference on distributed multi-modal data streams," in *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*, 2021, pp. 67–80.