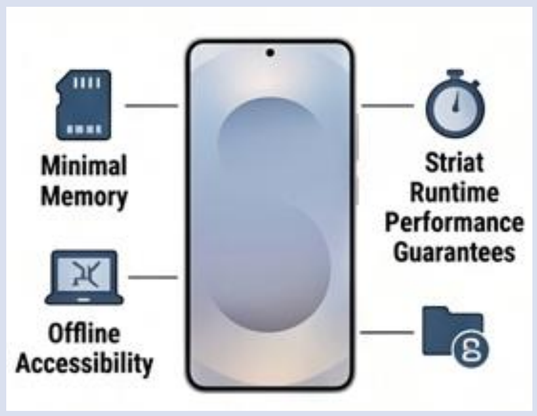
Unlocking the Edge deployment and ondevice acceleration of multi-LoRA enabled one-for-all foundational LLM

Sravanth Kodavanti†, Sowmya Vajrala†, Srinivas Miriyala†,
Utsav Tiwari†, Uttam Kumar†, Utkarsh Kumar Mahawar†, Achal Pratap Singh†, Arya D†, Narendra Mutyala†,
Vikram N R†, Sharan Allur†, Euntaik Lee‡, Dohyoung Kim‡, HyeonSu Lee‡, Gyusung Cho‡, JungBae Kim‡
† Samsung Research India, ‡ Samsung Electronics Korea

ACL 2026 SAN DIEGO

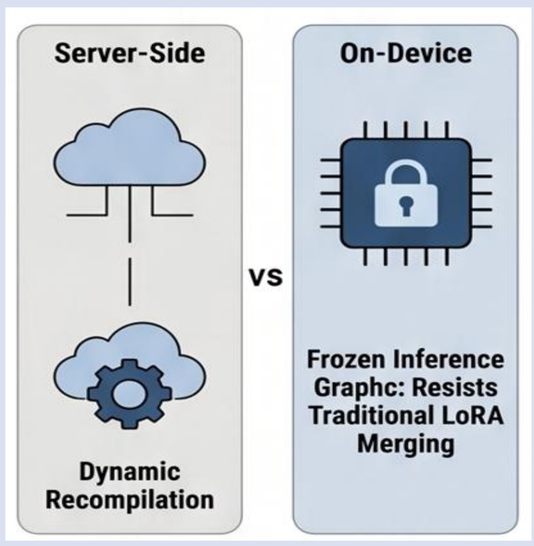
Problem & Motivation

Stringent Edge Constraints



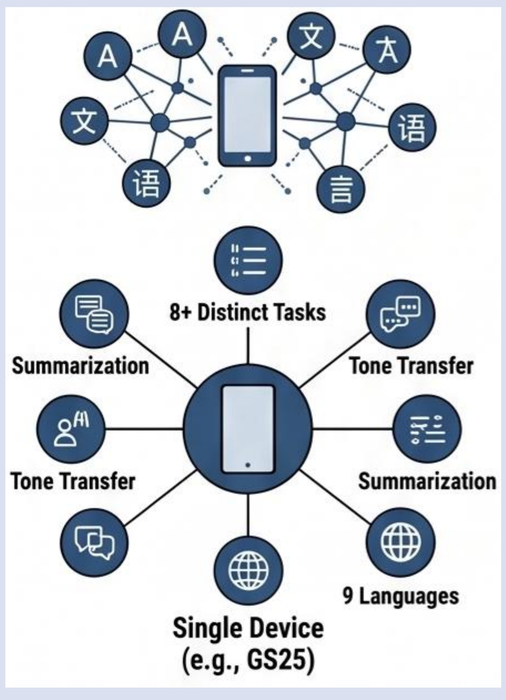
Mobile deployment is limited by minimal memory, strict runtime performance guarantees and the need for offline accessibility.

The Disconnect of Immutable Graphs



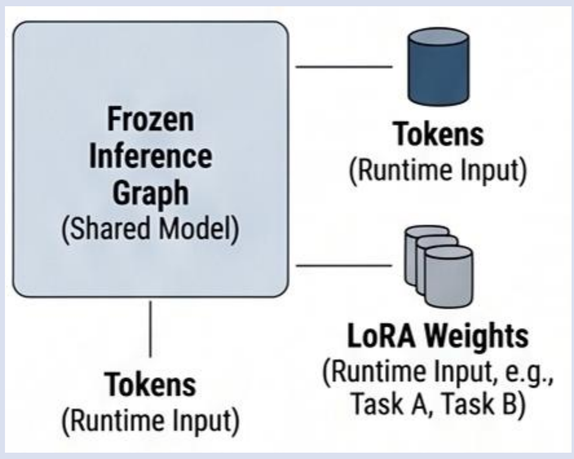
Server-side models can be dynamically recompiled , but on-device deployment requires frozen inference graphs that resists traditional LoRa merging.

One-for-All Multilingual Goal



Methodology - Optimizations

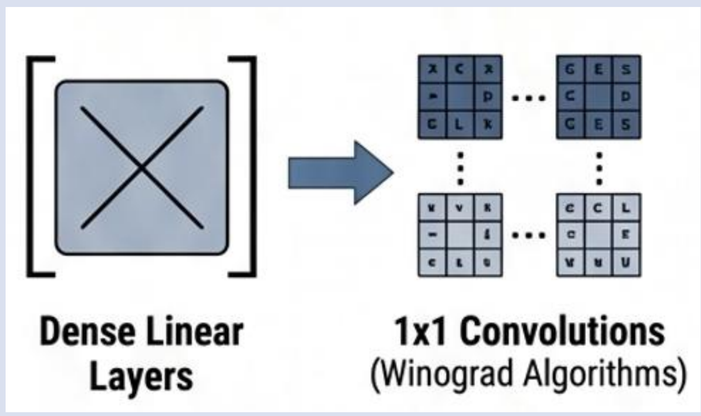
LoRA Weights as Runtime Inputs



Instead of static weight merging, we treat LoRAs as runtime inputs passed into placeholders within the frozen graph .

This enables "**plug-and-play**" task switching across **8+** tasks and **9** languages without memory overhead or the need to re-quantize the base model

Hardware aware Optimizations

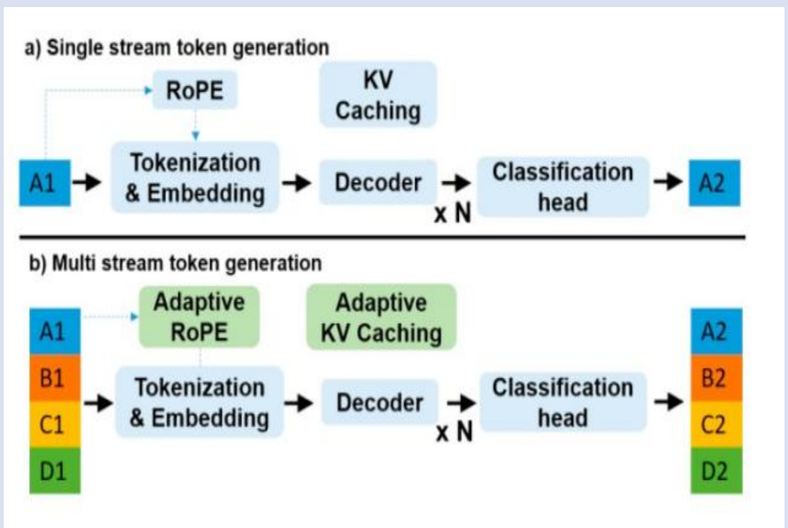


Attention Parallelism: MHA is decomposed into SHA to align with parallel NPU cores

Linear-to-Convolutional: Dense layers are reparameterized as 1x1 convolutions to exploit optimized Winograd kernels on mobile NPUs

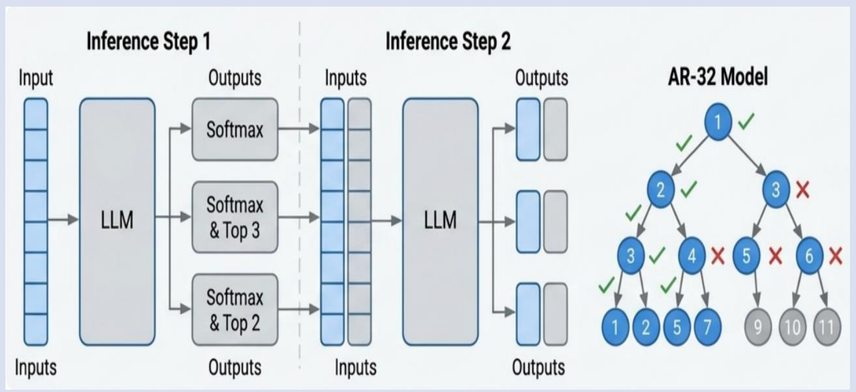
Innovative Decoding Strategies

Concurrent Token Generation (CTG)



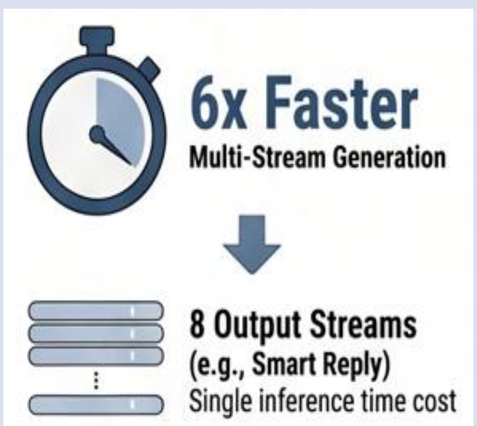
CTG generates up to **8 output streams** simultaneously over a shared KV-cache in a single forward pass.

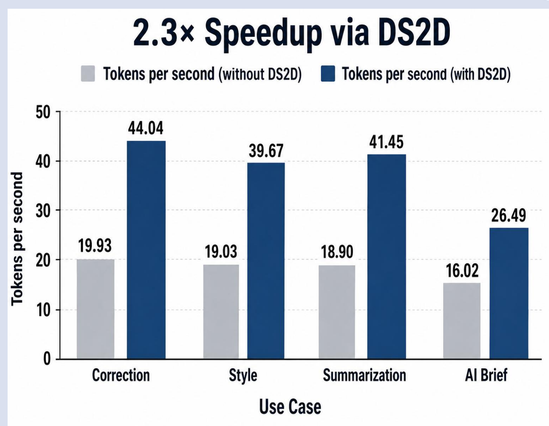
Dynamic Self Speculative Decoding (DS2D)



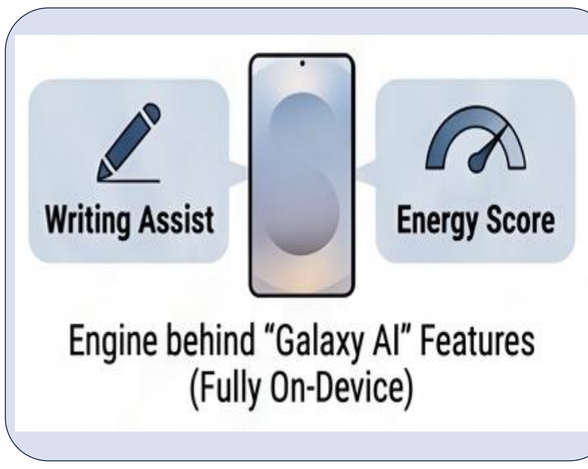
DS2D, Unlike standard speculative decoding, it requires no separate draft model, uses tree-based pruning with prefix tuning to predict future tokens, making it highly memory-efficient for mobile devices.

Key Results





Conclusions



Engine behind "Galaxy AI" Features (Fully On-Device)