Efficient On-Device Intelligence: From Continual Adaptation to Generative AI on Edge Devices

Young D. Kwon
Samsung AI Center-Cambridge, University of Cambridge
United Kingdom
yd.kwon@samsung.com

Abstract

Deploying capable AI on edge devices remains constrained by memory, compute, and energy. This paper presents a research programme across two fronts: (1) resource-efficient on-device training and adaptation that enable learning with minimal data without catastrophic forgetting; and (2) on-device generative AI that accelerates and compresses large-scale diffusion models and LLMs via position-aware compression and speculative execution. Through system-algorithm co-design, these techniques target real hardware constraints. Beyond this research, I have directly contributed to three consecutive on-device AI commercialisations on flagship Samsung smartphones, impacting millions of users.

CCS Concepts

• **Human-centered computing**; • **Computer systems organization** → *Embedded and cyber-physical systems*;

Keywords

Efficient ML, Generative AI, LLM Acceleration, Diffusion Model Compression, Continual Learning, Test-Time Adaptation, Speculative Decoding, Mobile Systems, Edge Computing, IoT

ACM Reference Format:
Young D. Kwon. 2026. Efficient On-Device Intelligence: From Continual Adaptation to Generative AI on Edge Devices. In *The 24th Annual International Conference on Mobile Systems, Applications and Services (MobiSys Companion '26), June 21–25, 2026, Cambridge, United Kingdom.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3812835.3814808

1 Introduction

Edge devices–from microcontrollers and mobile phones to embedded GPUs–are increasingly expected to perform sophisticated AI tasks entirely on-device, yet the gap between model demands and hardware capabilities remains vast. This tension manifests in two areas: (1) *on-device training and adaptation* (essential for personalisation for different users and changing environments), where standard pipelines assume resources that mobile and edge platforms cannot provide; and (2) *on-device generative AI*, where diffusion models and large language models (LLMs) demand scales that far exceed mobile and edge hardware budgets. My research addresses these challenges through system-algorithm co-design, developing solutions that target real hardware constraints. Figure 1 illustrates

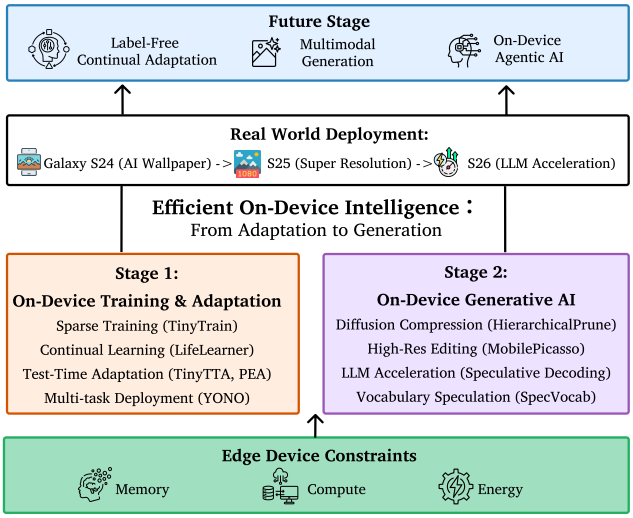

Figure 1: Overview of the research programme: From adaptation to generative AI on edge devices.

my research programme across past, present, and future. Beyond the research, I have directly contributed to the commercialisation of on-device AI technologies on Samsung Galaxy smartphones (S24, S25, S26, Z Flip6 & Fold6, Z Flip7 & Fold7), including personalised AI Wallpaper, efficient super-resolution, and on-device LLM acceleration, demonstrating that principled efficiency research can translate into products impacting millions of users.

2 On-Device Training and Adaptation

Enabling devices to learn and adapt locally is critical for personalisation and privacy-preserving deployment, yet edge hardware imposes constraints on memory, compute, and data availability.

Efficient Multi-Task Deployment and Sparse Training. At the most constrained end of the spectrum, YONO [9] enables multiple heterogeneous neural networks to share a single pair of codebooks, drastically reducing storage (by up to 12×) for multi-task deployment on microcontrollers. Also, UR2M [5, 6] brings uncertainty-aware event detection to microcontrollers. Scaling up to training on resource-constrained devices, TinyTrain [11] introduces a resource-aware sparse training framework that jointly captures user data, the memory, and the compute capabilities of the target device, drastically reducing backward-pass memory (1,098×) and computational cost (7.68×) while maintaining competitive accuracy.

Hardware-Aware Continual Learning. Empirical studies [7, 14] revealed that existing continual learning algorithms exhibit vastly different resource–accuracy trade-offs on real edge platforms

MobiSys Companion '26, Cambridge, United Kingdom

ACM ISBN 979-8-4007-2711-5/2026/08
https://doi.org/10.1145/3812835.3814808

than reported in algorithmic evaluations, motivating the hardware-aware designs that followed. LifeLearner [10] formulates continual learning as a hardware-aware meta-learning problem, learning an initialisation and update strategy that respects per-device memory and latency budgets while mitigating catastrophic forgetting. ContAuth [1] and FastICARL [8] apply continual learning to behavioural authentication and audio sensing, respectively, adapting and learning new tasks on resource-constrained devices without retraining from scratch.

Efficient Test-Time Adaptation. TinyTTA [4] makes test-time adaptation (TTA) feasible on edge devices through an early-exit ensemble strategy that selectively adapts lightweight auxiliary heads at intermediate layers, avoiding full-model backpropagation. PEA [18] removes the need for backpropagation entirely, aligning feature embeddings in a forward-pass-only manner using nearest-centroid alignment, making it architecture-agnostic and suitable for quantised on-device models [3]. Tempora [15] complements these by showing that the utility of online adaptation is highly time-contingent and that ignoring on-device resource constraints can significantly degrade performance [2, 19].

3 On-Device Generative AI

My recent work targets deploying large-scale generative models on edge devices, addressing both image generation via diffusion models and text generation via LLMs.

Diffusion Model Compression. HierarchicalPrune [13] introduces a position-aware structured pruning framework exploiting the observation that different layers in diffusion models contribute unequally to output quality. By assigning different pruning and preservation ratios per layer rather than a uniform pruning ratio or full model distillation, it removes up to 20-30% of parameters with minimal perceptual degradation. In addition, MobilePicasso [12] enables high-resolution image editing on-device by intelligently decomposing the task into low-resolution editing, learned projection from low-to high-resolution latents, and high-resolution denoising, avoiding hallucinations while achieving a drastic speed-up within on-device memory constraints.

LLM Acceleration via Speculative Decoding. Autoregressive LLM inference on edge devices is bottlenecked by memory bandwidth, leaving compute units underutilised. Speculative decoding addresses this by using a lightweight draft model to propose multiple candidate tokens that the target model verifies in a single batched forward pass, yielding wall-clock speedup while keeping the output distribution intact. However, as vocabulary sizes in modern LLMs continue to grow, computing the draft model's output distribution over the full vocabulary becomes the dominant cost during drafting. Unlike prior methods that use a fixed vocabulary subset and suffer rejections for out-of-subset tokens, SpecVocab [16] dynamically selects a contextually relevant subset at each decoding step, achieving higher acceptance lengths with over 15× smaller vocabulary size.

4 Commercialisation & Real-World Impact

Over three consecutive years, I have directly contributed to on-device AI commercialisations on flagship Samsung smartphones. In 2024, our team shipped *Personalised On-Device AI Wallpaper* on the Galaxy S24, Z Flip6, and Z Fold6, later extended to the S25, Z Flip7, and Z Fold7. In 2025, we delivered *Efficient On-Device Super Resolution* on the Galaxy S25 series. In 2026, we developed *Efficient On-Device LLM Acceleration via Speculative Decoding*, enabled on the Galaxy S26 series. Such real-world deployments demonstrate that the efficiency techniques in this programme can help meet commercial latency, memory, and power requirements.

5 Conclusion and Future Directions

This paper has presented a research programme spanning efficient on-device training and adaptation to on-device generative AI, grounded in system-algorithm co-design and demonstrated through three consecutive Samsung Galaxy commercialisations.

Looking ahead, three directions emerge. First, *label-free continual adaptation*: enabling robust and efficient continual adaptation of the deployed models without labels across diverse real-world scenarios, tasks, architectures, and applications. Second, *efficient multimodal generation*: bringing vision-language and multimodal models to edge platforms by co-designing visual and textual decoding pipelines that share computation across modalities. Third, *on-device agentic AI*: extending speculative decoding from general chat and multi-turn reasoning to agentic workflows, which consume substantially more tokens [17] and demand new efficiency strategies. These directions move toward a future where edge devices are not merely inference endpoints but continually evolving, multimodal, adaptive AI agents.

References

[1] Chauhan, Kwon, et al. 2020. ContAuth: Continual Learning Framework for Behavioral-Based User Authentication. *IMWUT* (2020).

[2] Danilowski, Murphy, Kwon, et al. 2026. MORPHEUS: Meta Test-Time Adaptation via Neural Collapse Geometry. In *ICLR 2026 TTU Workshop*.

[3] Dong, Kwon et al. 2025. LeanTTA: A Backpropagation-Free and Stateless Approach to Quantized Test-Time Adaptation on Edge Devices. arXiv:2503.15889

[4] Jia, Kwon, et al. 2024. TinyTTA: Efficient Test-time Adaptation via Early-exit Ensembles on Edge Devices. In *NeurIPS 2024*.

[5] Jia, Kwon et al. 2024. UR2M: Uncertainty and Resource-Aware Event Detection on Microcontrollers. In *2024 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. 1–10. doi:10.1109/PerCom59722.2024.10494467

[6] Jia, Kwon et al. 2026. A cascade framework for on-device uncertainty-aware event detection on microcontrollers. *Pervasive and Mobile Computing* 119 (2026), 102208. doi:10.1016/j.pmcj.2026.102208

[7] Kwon et al. 2021. Exploring System Performance of Continual Learning for Mobile and Embedded Sensing Applications. In *ACM/IEEE SEC'21*.

[8] Kwon et al. 2021. FastICARL: Fast Incremental Classifier and Representation Learning with Efficient Budget Allocation in Audio Sensing Applications. In *Proc. Interspeech 2021*. 356–360.

[9] Kwon et al. 2022. YONO: Modeling Multiple Heterogeneous Neural Networks on Microcontrollers. In *IPSN 2022*.

[10] Kwon et al. 2024. LifeLearner: Hardware-Aware Meta Continual Learning System for Embedded Computing Platforms. In *SenSys 2023*.

[11] Kwon et al. 2024. TinyTrain: Resource-Aware Task-Adaptive Sparse Training of DNNs at the Data-Scarce Edge. In *ICML 2024*.

[12] Kwon et al. 2025. Efficient High-Resolution Image Editing with Hallucination-Aware Loss and Adaptive Tiling. arXiv:2510.06295

[13] Kwon, Li, et al. 2026. HierarchicalPrune: Position-Aware Compression for Large-Scale Diffusion Models. In *AAAI 2026*.

[14] Li, Kwon et al. 2026. MetaCLBench: Meta Continual Learning Benchmark on Resource-Constrained Edge Devices. arXiv:2504.00174

[15] Sreeram, Kwon, et al. 2026. Tempora: Characterising the Time-Contingent Utility of Online Test-Time Adaptation. In *ICML 2026*.

[16] Williams, Kwon, et al. 2026. Speculative Decoding with a Speculative Vocabulary. arXiv:2602.13836

[17] Wu et al. 2025. Combating the Memory Walls: Optimization Pathways for Long-Context Agentic LLM Inference. arXiv:2509.09505

[18] Xiao, Kwon, et al. 2026. Architecture-Agnostic Test-Time Adaptation via Backprop-Free Embedding Alignment. In *ICLR 2026*.

[19] Xiao, Kwon, et al. 2026. Efficient Test-Time Adaptation via Decoupled BN Update For Edge Devices. In *ICLR 2026 TTU Workshop*.