# EMDL 2026: The Role of Mobile Computing in the Generative AI Era

7th International Workshop on Embedded and Mobile Deep Learning
**Theme: System Foundations for Generative AI at the Edge**
Co-located with ACM MobiSys 2026
Cambridge, United Kingdom – TBD, June 2026

**Overview:** Since its inception in 2017, the EMDL workshop has tracked how breakthroughs in deep learning transformed the interpretation of sensor data for mobile systems like smartphones and wearable devices. In the early years, the community focused on making standard inference feasible, overcoming the severe demands that deep models exerted on local resources. By 2022, these methods had matured, successfully adapting CNN and RNN architectures to meet the stringent needs of mixed-reality and cyber-physical systems.

**The Shift: From Discriminative to Generative:** However, the landscape has shifted once again. We are witnessing a transition from Discriminative AI (classifying sensor data) to Generative AI (reasoning, explaining, and acting on context). While Generative AI (GenAI) brings unprecedented capabilities, it also presents a resource wall. Modern edge devices operate under constraints in memory bandwidth and energy availability that standard GenAI architectures—which are memory-bound and autoregressive—fundamentally exceed. Currently, the most advanced models reside almost exclusively on the cloud, challenging the autonomy of mobile platforms.

In this context, the mobile computing community is in a unique position to begin the careful study of two core technical questions re-framed for the GenAI era. First, how should systems be architected to partition these massive workloads? We must move beyond simple offloading to explore dynamic collaboration where mobile devices handle context and lightweight generation while the cloud supports heavy lifting. Second, what is required to integrate GenAI into resource-constrained systems? This necessitates a re-examination of efficiency, spanning from the compression of Transformer architectures and diffusion models to the software/hardware optimization of mobile processors (CPUs, GPUs, NPUs) for memory-intensive generation rather than traditional convolution.

**Scope and Goals: EMDL 2026** explores the intersection of **Systems** and **Generative AI**. Unlike traditional AIoT approaches that focus on lightweight classification, this workshop addresses the unique systems challenges of GenAI deployment. We focus on the full stack of efficient deployment: from algorithmic compression to hardware-software co-design for resource-efficient reasoning on wearables, robots, and mobile devices. We invite researchers to submit work that answers core technical questions for the GenAI era:

1. **Architecture:** How should systems be architected to partition massive workloads between the cloud and the edge?.
2. **Efficiency:** What is required to integrate memory-bound GenAI into resource-constrained systems?.
3. **Edge-Native Design:** How do we define edge-native generative models designed explicitly for physical constraints?

## Topics of Interest

We solicit submissions including full technical workshop papers, white position papers, and work-in-progress/demos. Topics include, but are not limited to:

### Systems & Runtime

- **Runtime Systems:** Inference engines and runtime management specifically for GenAI.
- **Memory Optimization:** KV cache optimization and memory management for autoregressive generation.
- **Hardware Acceleration:** Heterogeneous computing (CPU/GPU/NPU) optimization and FPGA mapping for transformers and diffusion models.
- **Collaborative Intelligence:** Dynamic collaboration/offloading where mobile devices handle context while the cloud supports heavy lifting.
- **Cloud-level Intelligence:** Enabling cloud-level intelligence on mobile platforms via system architectures and techniques.

### Models & Algorithms

- **Efficient GenAI:** On-device Large Language Models (LLMs) via quantization, pruning, and distillation.
- **Generative Media:** Diffusion models for text-to-image and image-to-video generation on edge devices.
- **Multimodal Systems:** Generative systems combining vision, language, and audio inputs.
- **Retrieval-Augmented Generation (RAG):** Mobile-centric RAG architectures for on-device context retrieval.

### Applications & The Physical World

- **Agentic AI:** Reasoning and agentic systems operating on resource-constrained devices.
- **Benchmarking:** Evaluation methods and benchmarking for mobile GenAI performance.
- **Real-World Deployment:** Experiences and case studies of GenAI.

## Workshop Organizers

### PC CHAIRS

Young D. Kwon (*Samsung AI, UK*)
Stylianos I. Venieris (*Samsung AI, UK*)
Nicholas D. Lane (*University of Cambridge & FlowerLabs*)
Özlem Durmaz Incel (University of Twente, NL)
Dolly Sapra (University of Amsterdam, NL)
Guohao Lan (Delft University of Technology, NL)
Le Viet Duc (University of Twente)

### FULL PAPER SUBMISSIONS

Solicited submissions include both full technical workshop papers and white position papers. The maximum length of such submissions is 6 pages (excluding references). If accepted, papers will be published by ACM and appear in the ACM Digital Library.

**Submission Deadline: Thursday, April 9th, 2026, AoE**

### WORK-IN-PROGRESS AND DEMO SUBMISSIONS

Abstracts describing work-in-progress and demonstrations are also welcome and warmly encouraged. Submissions are limited to 2 pages (excluding references), and if accepted, will be included in the program as a short oral presentation – but will only be published on the workshop website (not the ACM DL). Deadlines for this informal track remain open even past the early registration deadline of MobiSys 2026; author notifications will be rolling (*i.e.* max of 4 days after submission) to enable early authors to take advantage of available discounts.