



NEURAL INFORMATION
PROCESSING SYSTEMS

Qualcomm

Depth through Recurrence: Towards Ultra-Efficient On-Device ASR

Chen Feng, Tianyi Xu, Ramchalam Kinattinkara Ramakrishnan, Yicheng Lin,
Shaojie Zhuo, Zhaocong Yuan, Chenzheng Su, Xiaopeng Zhang

Qualcomm AI Research · NeurIPS 2026

Deep ASR doesn't fit on the edge

GOAL

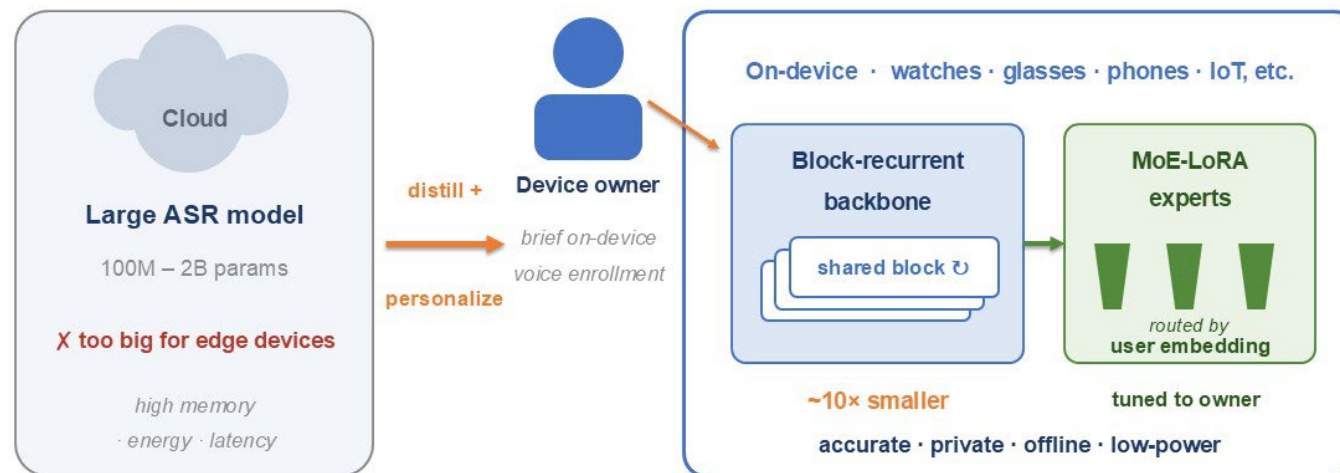
Accurate, private, always-on ASR on watches, glasses & IoT.

CHALLENGE

SOTA models have 100M–2B params; quantization / pruning alone is not enough.

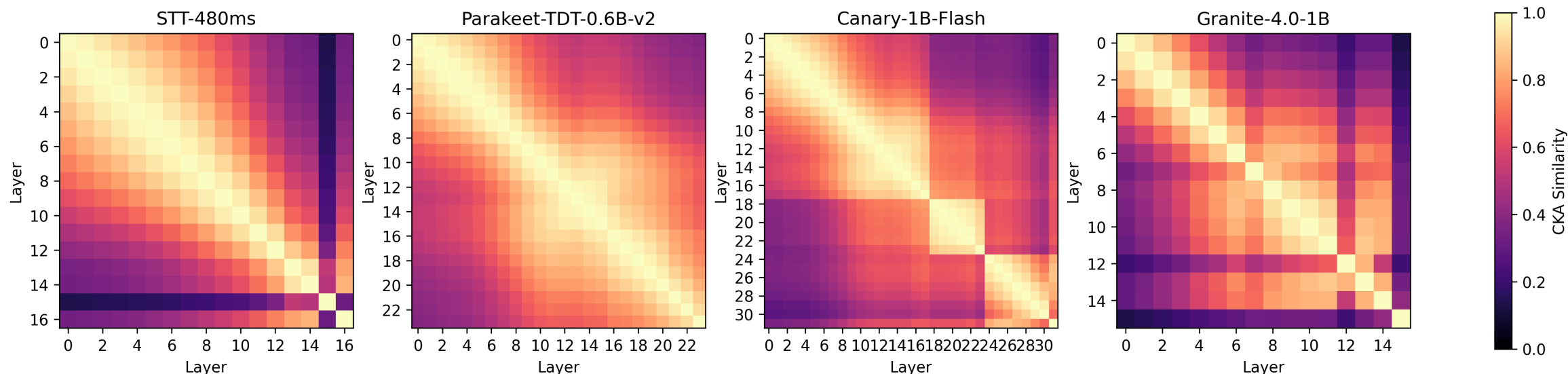
QUESTION

Does every encoder layer learn something unique?



100M – 2B parameters → megabyte-scale memory budget

Adjacent layers are similar



Layer-wise CKA similarity across four SOTA ASR encoders (STT-480ms, Parakeet-0.6B, Canary-1B, Granite-1B).

OBSERVATION

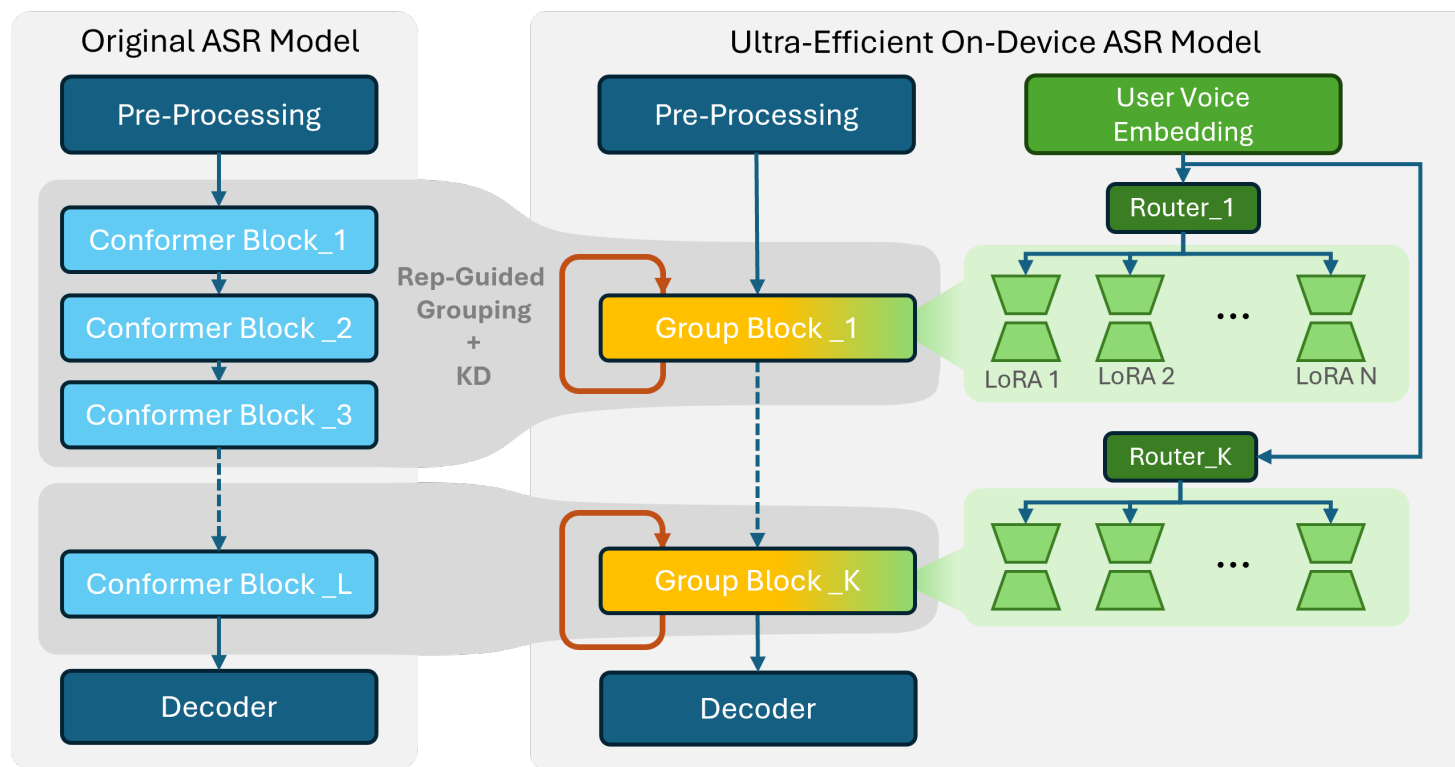
Strong block-diagonal structure: neighbouring layers learn functionally similar transformations.

BLOCK-RECURRENT HYPOTHESIS

$$\mathbb{E}_{x \sim \mathcal{D}} [\|f_L(x) - (B_K^{n_K} \circ \dots \circ B_1^{n_1})(x)\|_2] \leq \varepsilon$$

L layers $\approx$ $K \ll L$ shared blocks, each applied n_k times
 $\sum n_k = L$.

Method: group, distill, personalize



1

Group

CKA + DP $\rightarrow$ K contiguous groups

2

Initialize

average the teacher weights within each group

3

Distill

task + encoder KD + decoder KD

4

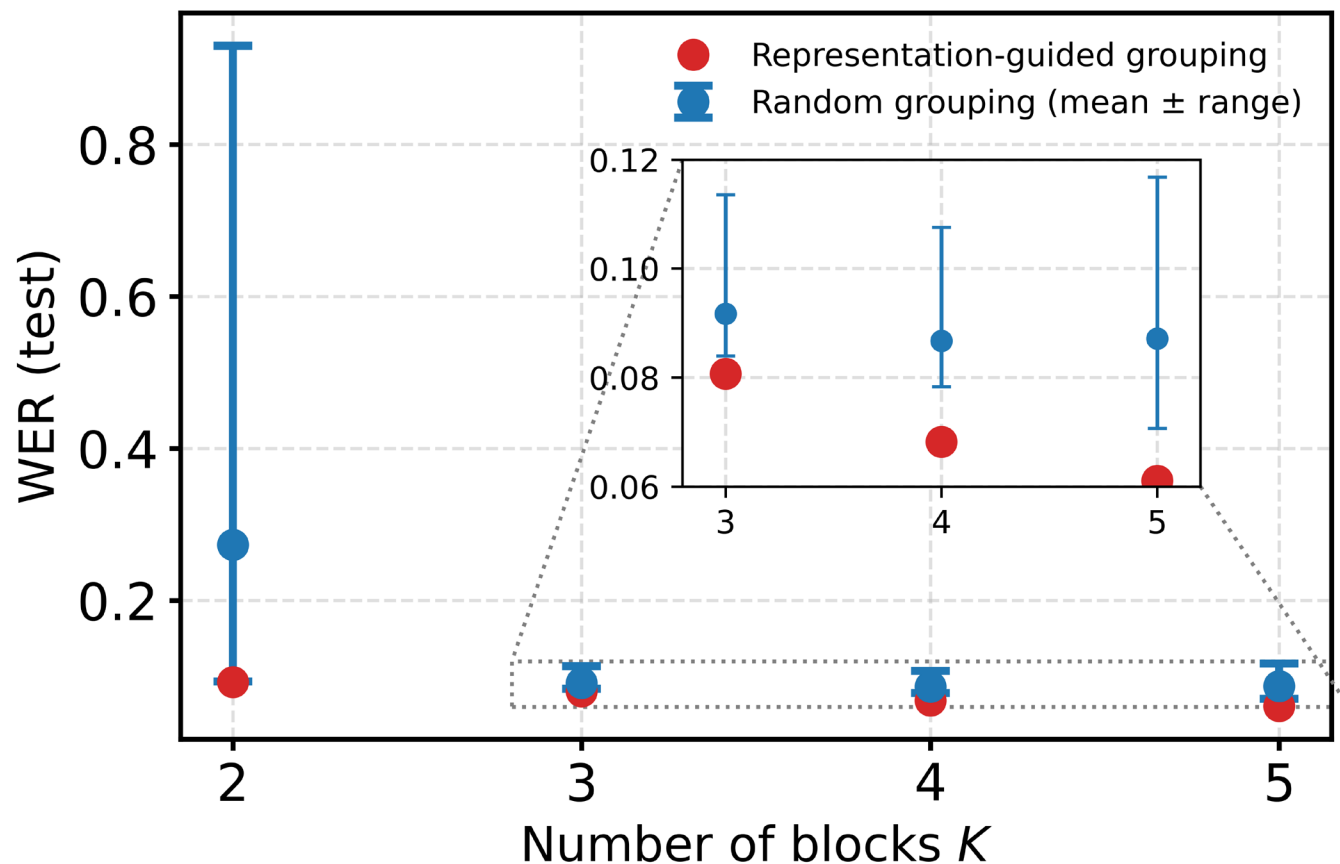
Personalize

user embedding routes MoE-LoRA experts

TRAINING LOSS

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{task}}(\theta) + \lambda_{\text{enc}} \mathcal{L}_{\text{KD}}^{\text{enc}}(\theta) + \lambda_{\text{dec}} \mathcal{L}_{\text{KD}}^{\text{dec}}(\theta)$$

Representation-guided grouping is critical



CKA-guided vs. random grouping — WER (test) by number of blocks K .

WHY IT MATTERS

At $K=2$, random grouping can collapse (WER up to ~ 0.9); guided grouping stays ~ 0.08 .

K IS THE DIAL

Larger $K \rightarrow$ higher accuracy. Smaller $K \rightarrow$ smaller footprint.

BONUS

Weight reuse improves cache locality: lower energy & latency at equal FLOPs.

Near-teacher accuracy, 7.7× smaller

Model	K	Params (M)	Avg WER %	Δ
STT-480ms (teacher)	—	108.9	11.05	—
Edge-STT	2	14.2	13.40	+2.35
Edge-STT	4	26.8	10.72	−0.33
Edge-STT	8	52.1	8.88	−2.17
Parakeet-v2 (teacher)	—	608.9	6.05	—
Edge-Parakeet	8	205	7.07	+1.02

Average WER over 8 Open ASR Leaderboard benchmarks.
K=2 keeps 97.4% of teacher accuracy; K=4 matches the teacher.

97.4%
of teacher accuracy at K=2

7.7×
smaller encoder (109M → 14M)

1.46×
less memory energy per inference

Personalization helps every speaker

1 ENROLL

short on-device voice embedding

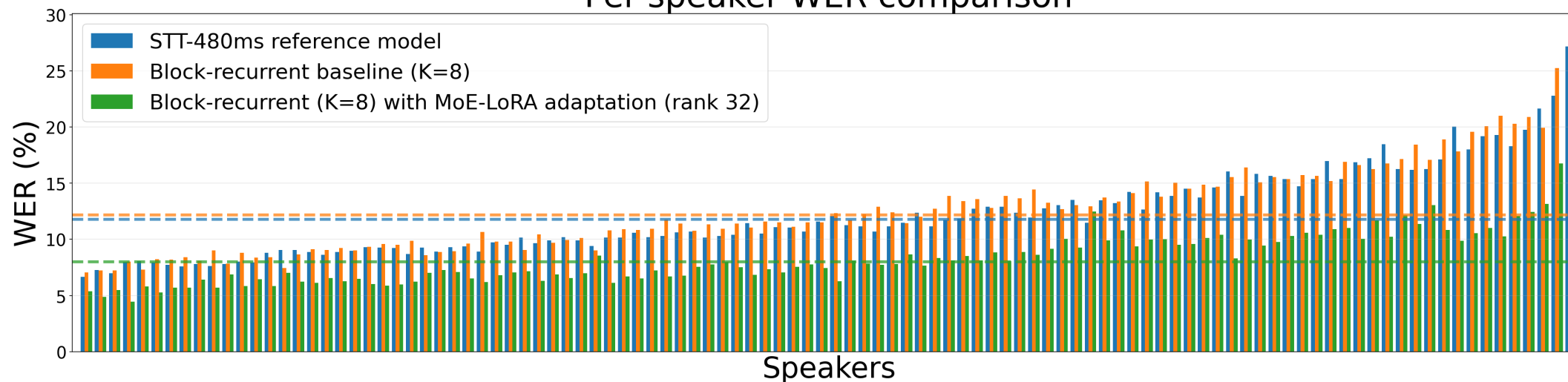
2 ROUTE

top-k MoE-LoRA experts per block

3 FUSE

frozen backbone, no extra latency

Per-speaker WER comparison

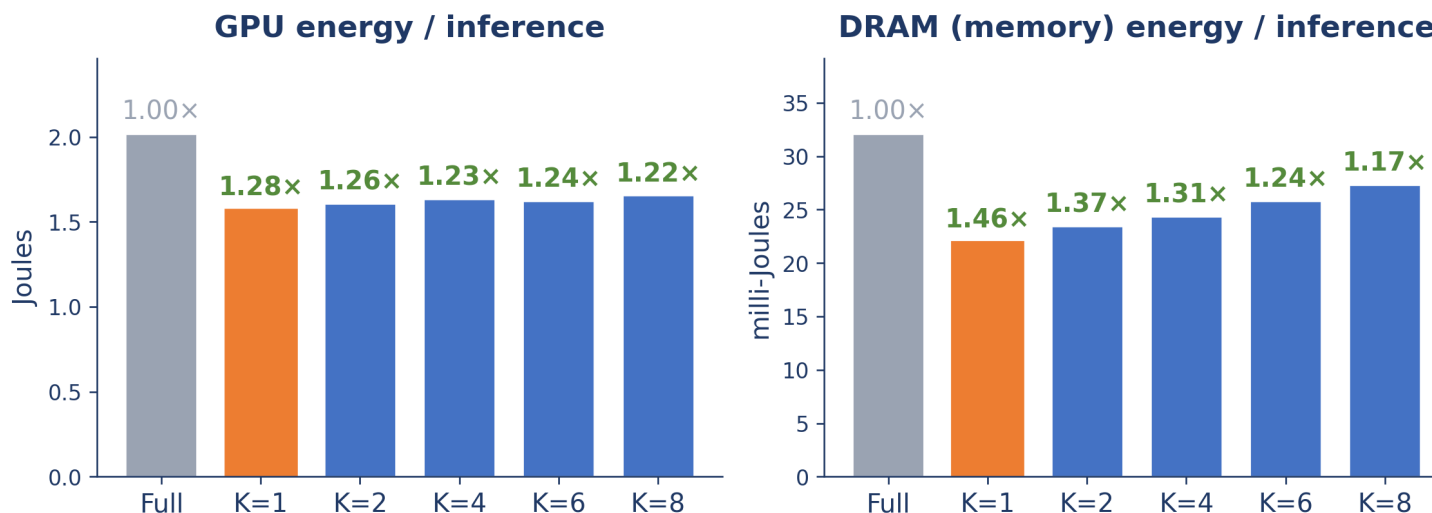


Model	K	Zero-shot	r16	r32	r64
Edge-STT	8	12.19	8.18	8.03	8.04
Edge-Parakeet	8	8.12	5.24	5.20	5.10

WER improves for all 106 speakers

Weight reuse cuts energy and latency

Up to 1.46x energy reduction



Memory access costs far more energy than compute — and today's hardware under-exploits weight reuse, so there is headroom for co-design.

1.46x

lower DRAM energy

1.28x

lower GPU energy / inference

1.21x

lower latency at comparable FLOPs

Depth ≠ parameters: structured recurrence → compact, accurate, efficient, personalized ASR



NEURAL INFORMATION
PROCESSING SYSTEMS

Qualcomm

Thank you!

Questions welcome

Depth through Recurrence: Towards Ultra-Efficient On-Device ASR

Qualcomm AI Research · NeurIPS 2026 · chenf@qti.qualcomm.com