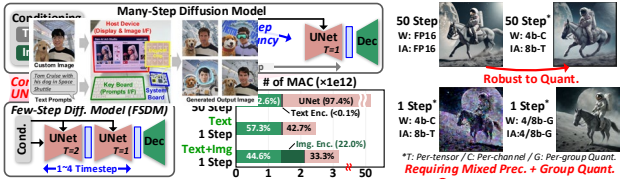
EdgeDiff: Multi-modal Few-step Diffusion Model Accelerator with Mixed-Precision and Reordered Group-Quantization for On-device Generative AI

Sangjin Kim, Jungjun Oh, Jeonggyu So, Yuseon Choi, Sangyeob Kim, Dongseok Im, Gwangtae Park, and Hoi-Jun Yoo

KAIST

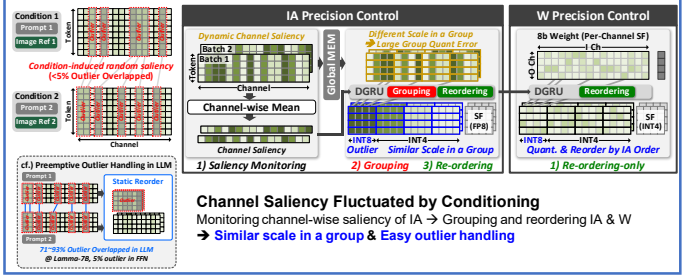
Motivation

Trending Algorithm Opt. for DM: Few-Step Diffusion Model with Knowledge Distillation



EdgeDiff: Accelerate All **Encoder/Unet/Decoder** of FSDM with **Mixed Precision + Group Quant.**

Proposed Condition-aware Reordered Mixed Precision



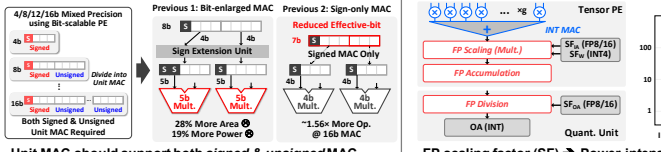
Channel Saliency Fluctuated by Conditioning

Monitoring channel-wise saliency of IA → Grouping and reordering IA & W

→ **Similar scale in a group** & **Easy outlier handling**

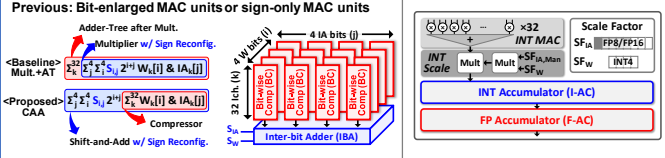
Proposed HW Opt. for *Mixed Precision* and *Group Quantization*

Micro-arch. for Mixed Precision MAC



Unit MAC should support both *signed & unsigned* MAC

Previous: Bit-enlarged MAC units or sign-only MAC units



Reducing Overhead by Changing Accum. Order

Baseline: **Inter-bit** (mult. w/ sign reconfig. ☹) → **Inter-ch.** (add)

CAA: **Inter-ch.** (compressor) → **Inter-bit** (add w/ sign reconfig. ☺)

Micro-arch. for Group Quantization

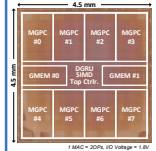
FP scaling factor (SF) → Power-intensive FP operation

Tiering into INT-Accum. & FP-Accum.

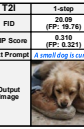
(Tier 1: I-AC) Checking Range & In-range Data Accumulation

(Tier 2: F-AC) Conditional Operation for Power Reduction

Implementation Results (Chip, System)



Evaluation on Public Dataset (T2I, I2I)

T2I	1-step	4-step
FID	20.09 (FP: 19.76)	19.25 (FP: 19.00)
CLIP Score	0.310 (FP: 0.321)	0.310 (FP: 0.317)
Text Prompt	*A small dog is curled up on top of the shoes*	
Output Image		

I2I	Prompt 1	Prompt 2	Prompt 3
Input Image		Reference Image from MS-COCO Dataset	
Text Prompt	*Pixar, 8k, Smiling*	*Minecraft*	*Sunglass, Cartoon*
Output Image			
PSNR	30.62 dB	31.19 dB	31.24 dB

Demonstration System with Custom Image