KYUNG HEE UNIVERSITY

Efficient Multimodal Intelligence Lab EMIL

KV Cache Quantization

Dongwon Lee
April, 14

Last Week

Action Items

- **Last Meeting Summary**
  - A White Paper on Neural Network Quantization (~Range Setting), PTQ~
  - SnapKV
- **Action Items**
  - A White Paper on Neural Network Quantization
  - **Multimodal** KV cache **quantization** Paper

Agenda

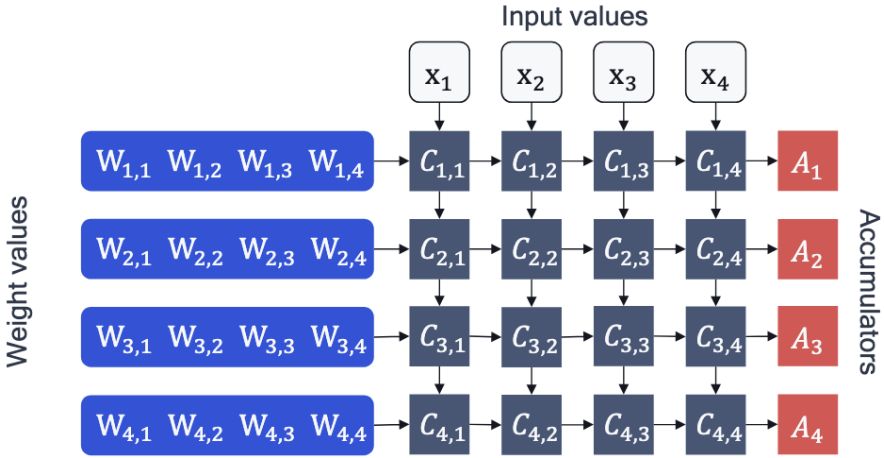
- Post Training Quantization(PTQ)
- Multimodal KV Cache Quantization

Post Training Quantization(PTQ)

Quantization Fundamentals

Quantization, Round-to-nearest

Matrix-multiply logic in neural network accelerator hardware



Uniform affine Quantization

Mapped to discrete integer

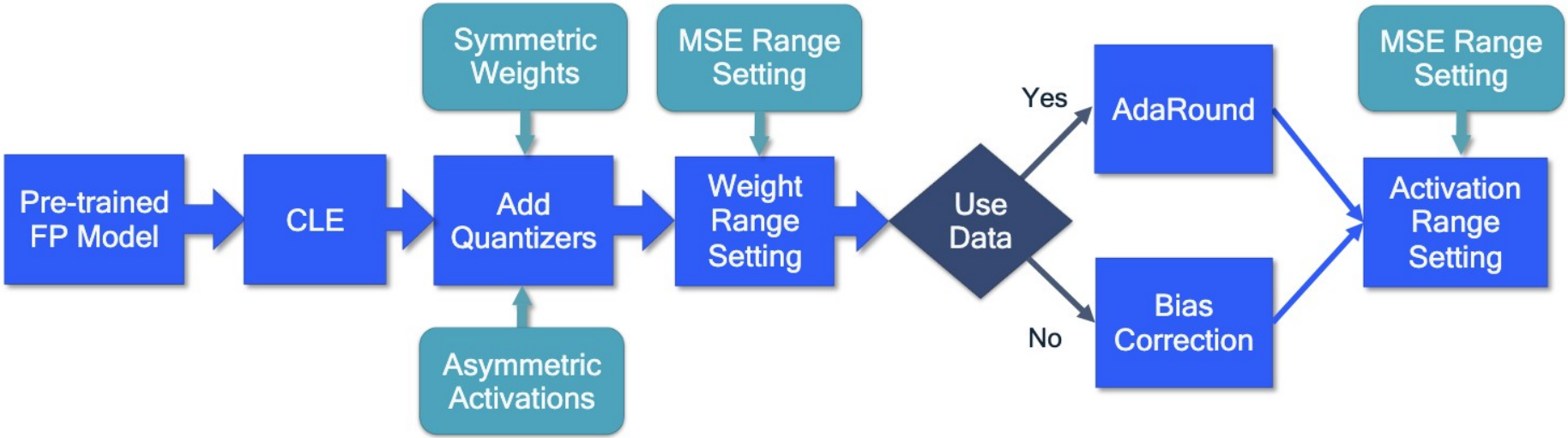
$$x_{\text{dis}} = \left\lfloor (x - \alpha) \cdot \frac{2^b - 1}{\beta - \alpha} \right\rceil$$

Dequantization

$$x_{\text{deq}} = x_{\text{dis}} \cdot \frac{\beta - \alpha}{2^b - 1} + \alpha.$$

Post Training Quantization(PTQ)

Pipeline



Cross-Linear Equalization (CLE)

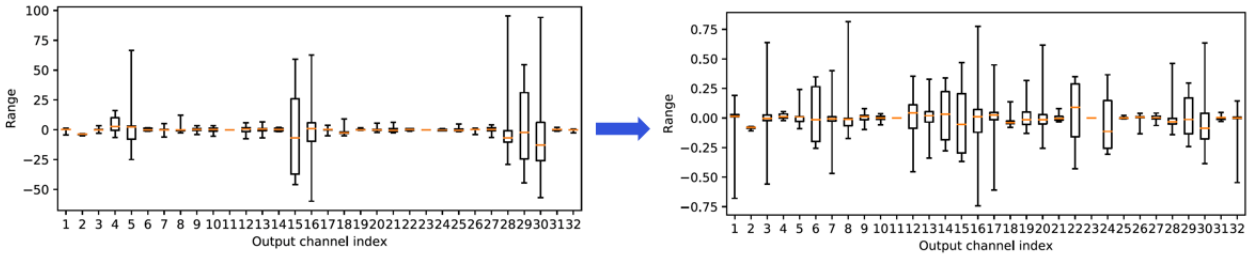
Post Training Quantization(PTQ)

- **Goal**: Better **weight** distribution **before quantization**
  - Reduced dynamic range
- **Method**: Reshaping Weight Distribution
  - Equalizing **dynamic ranges** across consecutive layers

$$\mathbf{y} = f(\mathbf{W}^{(2)}\mathbf{S}\hat{f}(\mathbf{S}^{-1}\mathbf{W}^{(1)}\mathbf{x} + \mathbf{S}^{-1}\mathbf{b}^{(1)}) + \mathbf{b}^{(2)})$$

$$\mathbf{y} = \mathbf{W}^{(2)}(f(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}) + \mathbf{c} - \mathbf{c}) + \mathbf{b}^{(2)}$$

Cross-Layer Equalization improves performance of Quantized models



Pre-equalization box chart

Post-equalization box chart

This equalization reduces the differences in dynamic-ranges across channels -> better quantization

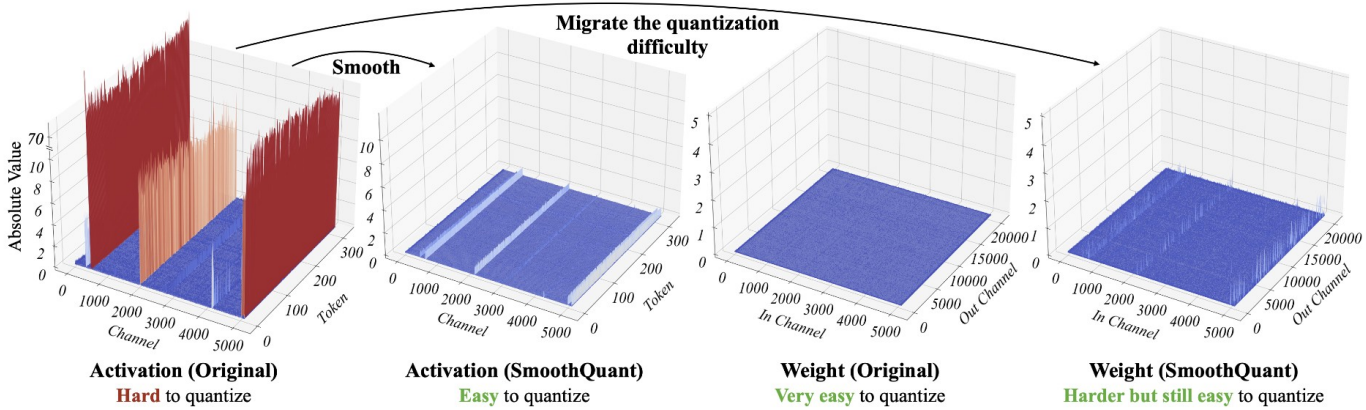
Qualcomm AIMET(AI Model Efficiency Toolkit) Docs

SmoothQuant (W8A8)

SmoothQuant: Accurate and Efficient Post-Training Quantization for Large Language Models

- **Problem:**
  - **Activation Outliers -↓ quantization error**
  - **per-channel is infeasible, system implementation difficult**
- **Method:** Activation-weight equalization (Migrating Difficulty from Activations to Weights)
- **Result:** Better distribution for **activation quantization**
- **Limitation: grid search a(hyperparameter), outlier is not solved(just migrated)**

$$\mathbf{Y} = (\mathbf{X}\mathrm{diag}(\mathbf{s})^{-1}) \cdot (\mathrm{diag}(\mathbf{s})\mathbf{W}) = \hat{\mathbf{X}}\hat{\mathbf{W}}$$



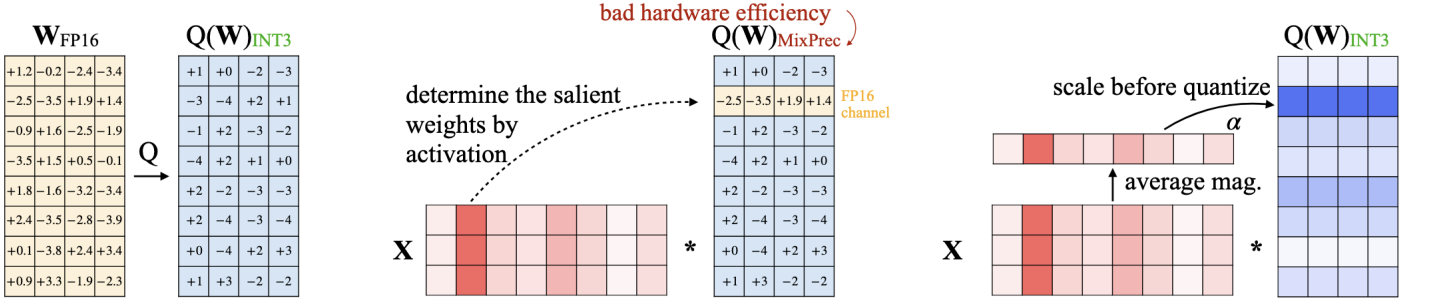
AWQ for Low-bit Weight-only Quantization

AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration

- **Goal:** mitigate weight quantization error (W4A16)
- **Idea:** weights are not equally important, larger activations are generally more important
- **Method:** Activation-aware weight protection
  - Channel-wise, **scale up salient weight** channels

$$Q(w \cdot s) \cdot \frac{x}{s} = \Delta' \cdot \text{Round}\left(\frac{ws}{\Delta'}\right) \cdot x \cdot \frac{1}{s},$$

- **Effect:**
  - Robust to calibration set size and domain shift
    - **Prevents over-fittingl — grid search over α**
  - **No backpropagation, no weight update — grid search over α**
    - **Generalize to different domains and modalities**
- **Limitation: weight-only, not activation**

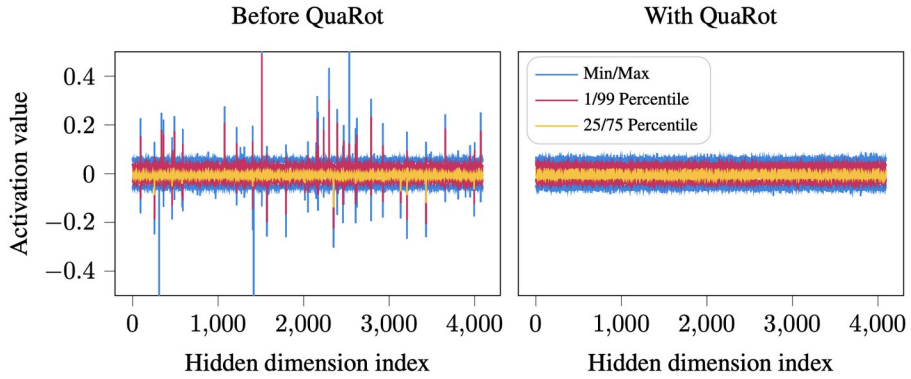


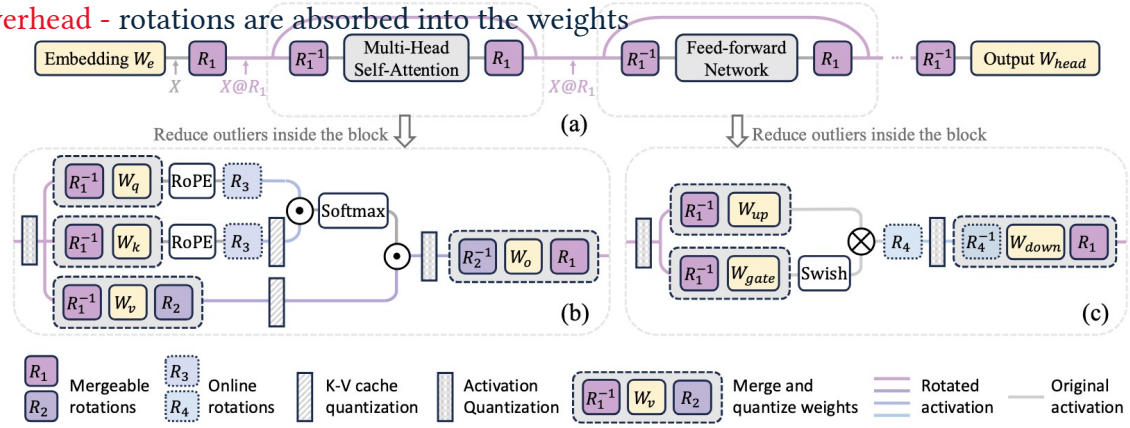
(a) RTN quantization (**PPL 43.2**) (b) Keep 1% salient weights in FP16 (**PPL 13.0**) (c) Scale the weights before quantization (**PPL 13.0**)

QuaRot / SpinQuant

Quarot: Outlier-free 4-bit inference in rotated llms / SpinQuant: LLM quantization with learned rotations

- **Goal:** Mitigate activation outliers for low-bit quantization (W4A4)
- **Method:** Rotation-based quantization
  - QuaRot: Hadamard rotation
  - SpinQuant: Learn the optimal rotation on the Stiefel manifold
- **Effect:**
  - Removes the **channel-aligned outlier**
  - Enables Outlier-free 4-bit inference
  - Preserves accuracy under low-bit quantization
  - no runtime overhead - rotations are absorbed into the weights



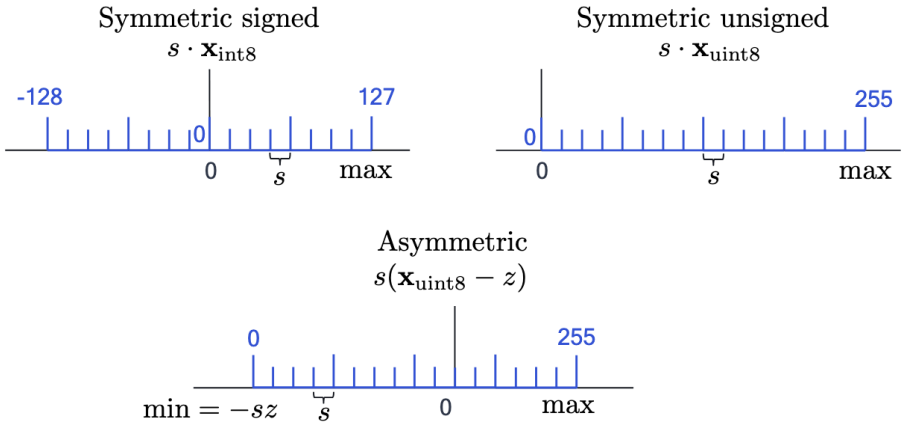


Symmetric v.s. Asymmetric

Quantizaer

- **Symmetric Quntization**
  - Weights are zero-centered

- **Asymmetric Quntization**
  - Activations are not necessarily zero-centered(activation function)
  - Common for activations in CNNs
  - Most KV Cache(KIVI, CalibQuant, VidKV, ...
    - Exceptional**:** 1bit

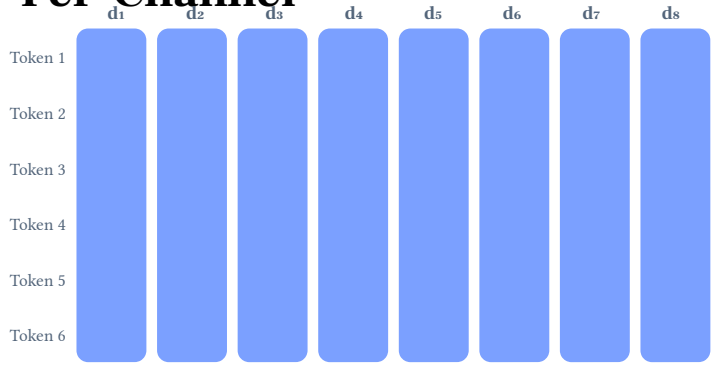


Granularity

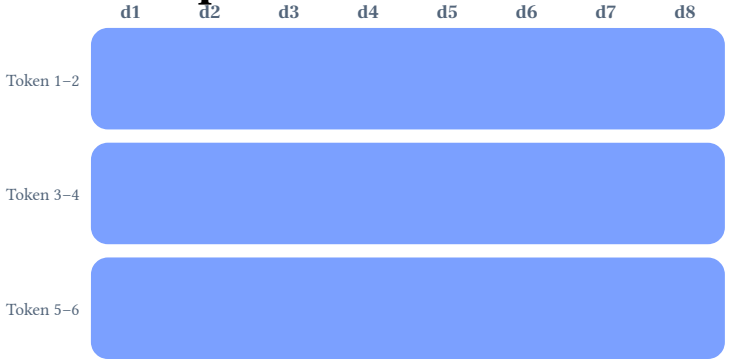
Quantization

- **Per-Tensor**

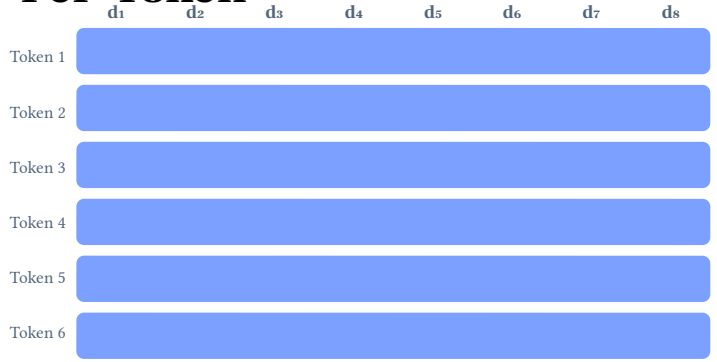
- **Per-Channel**



- **Per-Group**



- **Per-Token**



Range Setting

Post Training Quantization(PTQ)

Min–Max

$$\Delta = (x_{max} - x_{min}) / (2^b - 1)$$

$$z = x_{min}$$

$$q = \text{round}((x - z) / \Delta)$$

PROS
- No clipping error
- Simple and efficient

CONS
- Highly sensitive to outliers

MSE-based Range

$$(q_{min}, q_{max}) = \underset{q_{min}, q_{max}}{\arg\min} \left\| x - \hat{x} \right\|_2^2$$

Select the clipping range that minimizes reconstruction error

PROS
- Reduces the impact of outliers
- Better reconstruction quality

CONS
- Introduces clipping
- Activation range is data-dependent

Clipping Error ↔ Rounding Error

Bias Correction

Post Training Quantization(PTQ)

- **Problem:** Quantization Error(rounding error/clipping error) -> Mean Shift

$$\mathbb{E}\left[\mathbf{W}\mathbf{x}\right] \neq \mathbb{E}\left[\widehat{\mathbf{W}}\mathbf{x}\right]$$

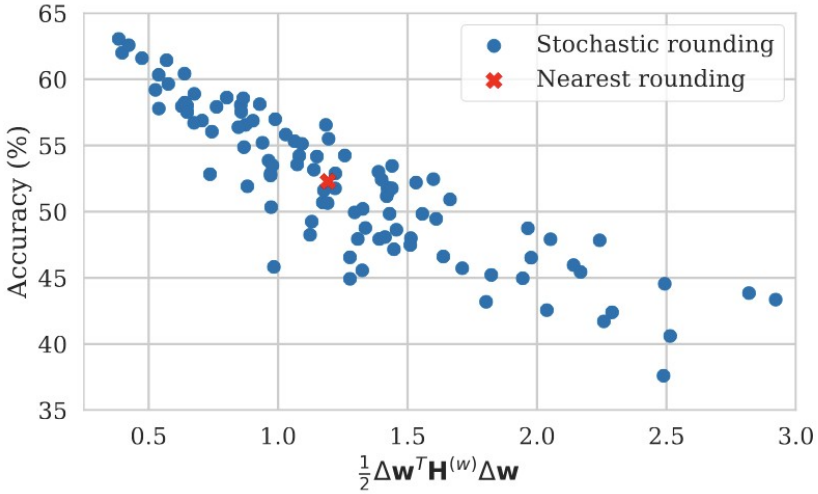
$$\mathbb{E}\left[\mathbf{y}_{\text{corr}}\right] = \mathbb{E}\left[\widehat{\mathbf{W}}\mathbf{x}\right] - \Delta\mathbf{W}\,\mathbb{E}\left[\mathbf{x}\right] = \mathbb{E}\left[\mathbf{y}\right]$$

- **Method**
  - Analytic (BN statistics) vs. empirical (**calibration data**)
  - Essentially zero cost, non-trivial gain — the cheapest compensation available

AdaRound

Post Training Quantization(PTQ)

- **Problem: round-to-nearest is not optimal**
- **Method:** Learns whether to round each weight up or down using calibration set
  - Layer-wise optimization
  - Optimize a reconstruction loss, no backpropagation



Comparison

PTQ - White Paper v.s. LLM

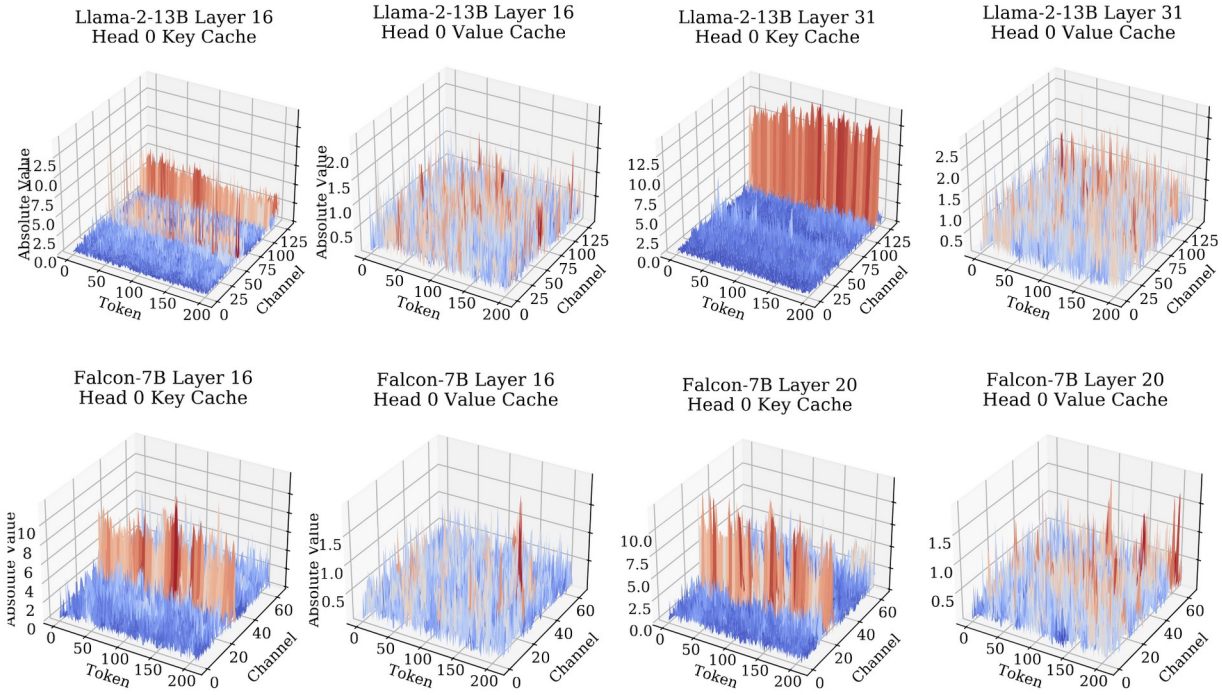
White Paper	LLM
Cross-Layer Equalization(CLE)	SmoothQuant, AWQ, SpinQuant, QuaRot
Range Setting	-
Bias Correction	-
Adaround	GPTQ
-	**KV Cache** - KIVI

KV Cache Quantization

KIVI

KIVI: A Tuning-Free Asymmetric 2bit Quantization for KV Cache

- Previous works:
  - lack of exploring **element distribution of KV cache**
- Method:
  - Key: **per-channel**, few fixed channels whose magnitudes are very large
  - Value: **per-token**. no obvious outlier pattern -> quantized per token
- Results: Peak memory, batch size, throughput
- Limitation: group scale overhead



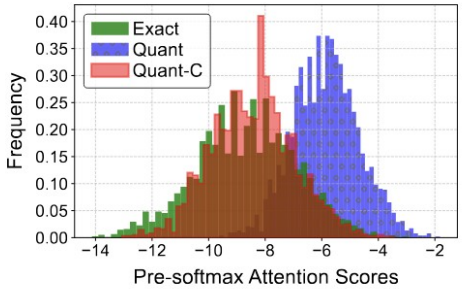
CalibQuant

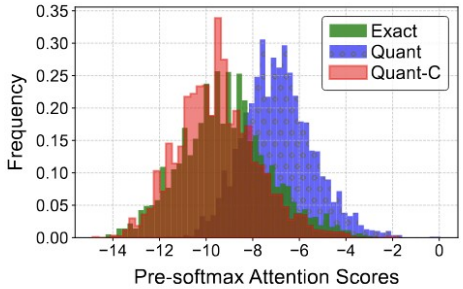
CalibQuant: 1-Bit KV Cache Quantization for Multimodal LLMs

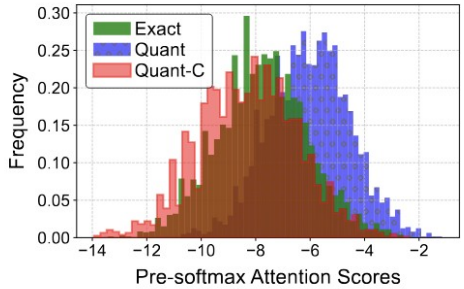
$$q \cdot k_{\text{deq}} = q \cdot \left(k_{\text{dis}} \odot \frac{\beta - \alpha}{2^b - 1} + \alpha \right)$$
$$= \left(q \odot \frac{\beta - \alpha}{2^b - 1} \right) \cdot k_{\text{dis}} + q \cdot \alpha$$

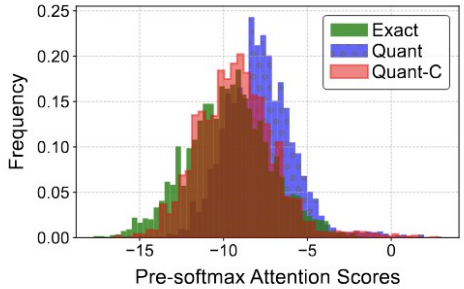
- Previous Works: general LLM
  -> MLLM visual token

- Method:
  - **visual** KV cache low-bit quantization (1, 2 bit)
  - **Channel-wise**: Key - per-channel, Value - per-channel
  - Post-scaling: decode 시 q는 1토큰 → dequant를 query 쪽으로
  - Calibration: 1-bit면 값이 $\{\alpha, \beta\}$으로 인한 pre-softmax 분포 왜곡 -> affine map $g: [\gamma, \delta] \rightarrow [\gamma, -\tau_1\delta - \tau_2]$, $\tau \in \{0, 1, 2, 3\}$ grid search





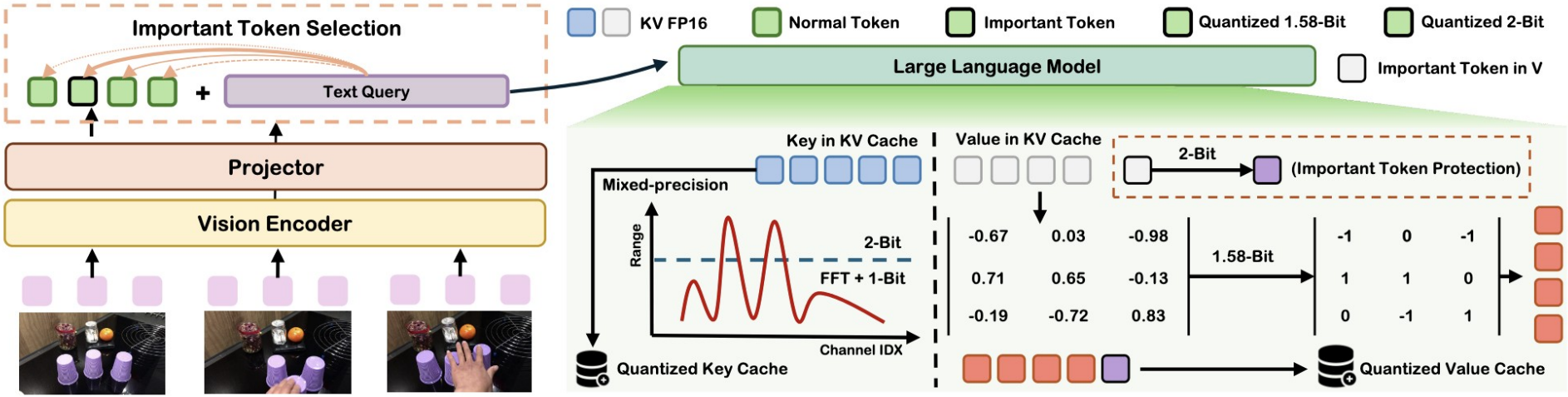




VidKV

PLUG-AND-PLAY 1.x-Bit KV CACHE QUANTIZATION FOR VIDEO LARGE LANGUAGE MODELS

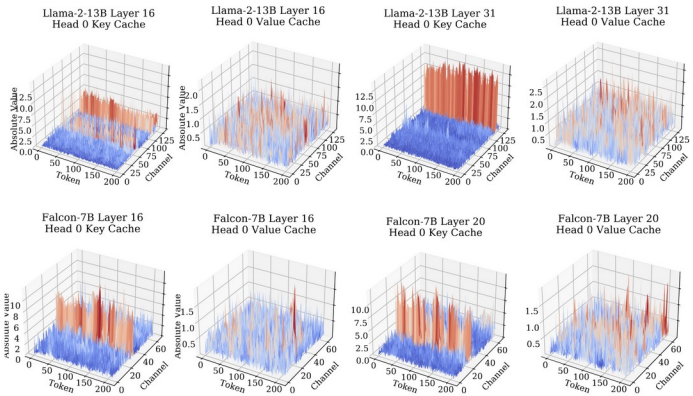
- separate bit budgets for keys and values in video LLMs, importance-driven mixed precision
- Idea: **High redundancy of video tokens** -> low-bit quantization
- Method:
  - Key: **Mixed precision** per channel(**2-bit**: anormal channel, **1-bit**: normal)
  - Value: **1.58-bit** quantization(mapping {−1, 0, 1} based on a & selectively filtering semantically salient visual tokens(import token – 2bit)
    - per-channel(<-> KIVI: per-token)



Discussion

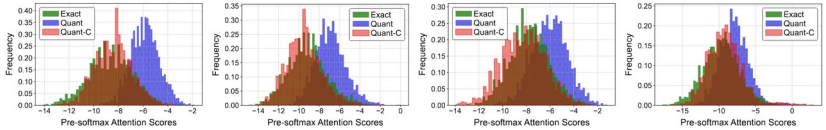
Plan

- KV Cache Quantization llm saturated .
  -
    - MHA(multi-head-attention) -> KV Cache DeepSeek-V2
  - LLM KIVI . (2024~)
    - K, V channel-wise
  - HW: NVFP4



- Multimodal or Edge ...

- Track 1: ???-aware quantization(e.g. modality-aware quantization, video distribution/statistical patterns) -> granularity , saliency
  - VidKV: Text, Image
  - KIVI: K, V granularity quantize
  - visualization -> modality aware quantization
    - KIVI K, V element distribution ...



- Track 2: kernel, channel-wise
  Granularity , salient channel kernel .

Thank You!