AAAI-25 / IAAI-25 / EAAI-25
FEBRUARY 25 – MARCH 4, 2025 | PHILADELPHIA, USA

What **Diffusion Model** weight layers are **sensitive** to **quantization?**

Which ones are **responsible** for inhibiting **low-bit quantization?**

Insights from a **myriad** of **models.**

UNIVERSITY OF ALBERTA

HUAWEI
ALBERTA INNOVATES

Qua²SeDiMo: Quantifiable Quantization Sensitivity of Diffusion Models

Keith G. Mills[1], Mohammad Salameh[2], Ruichen Chen[1], Negar Hassanpour[2], Wei Lu[3] and Di Niu[1]

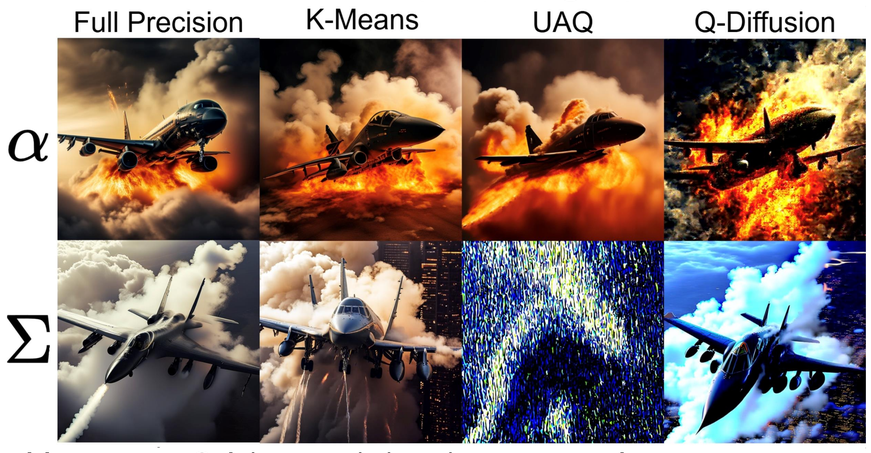
[1]Dept. ECE, University of Alberta [2]Huawei Technologies Canada [3]Huawei Kirin Solution

Diffusion Model Post-Training Quantization

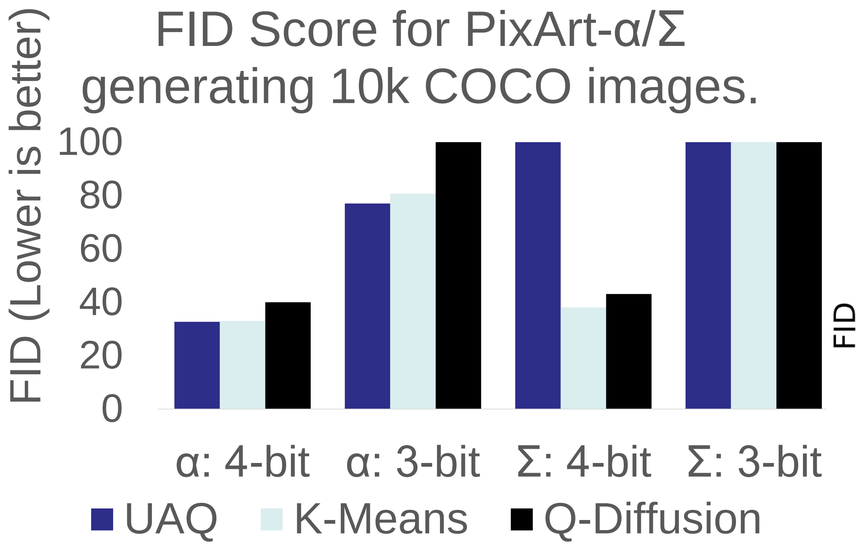
- Reduce DNN bit-precision.
  - E.g., from 16-bit to 8/6/4/3/-bit
- Weight and activation quantization.
- Post-Training Quantization (PTQ) is low-cost, doable on *already trained* DNN models.

Challenge: <4-bit Weight Quantization

Weight precision controls model size. Most methods achieve 4-bit (W4) weight precision:



However, 3-bit precision is a struggle.

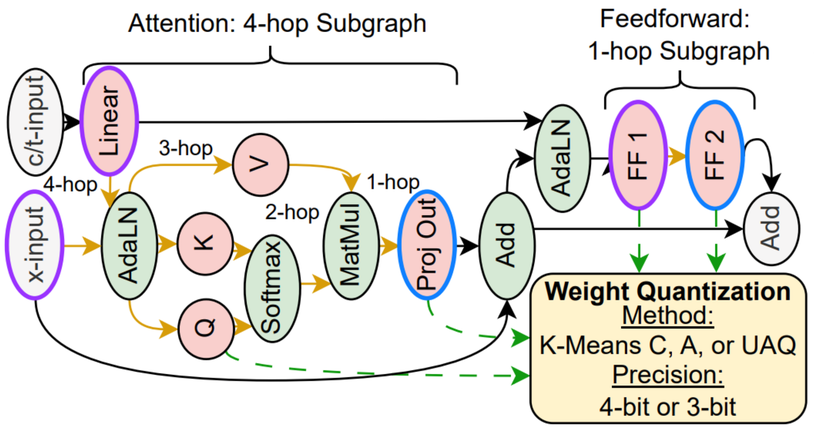


Our Hypothesis: Weight Sensitivity

Performance degradation caused by *some* weight types being quantized to 3-bit precision. Meanwhile, other weights are less crucial and can be lowered to 3-bits.

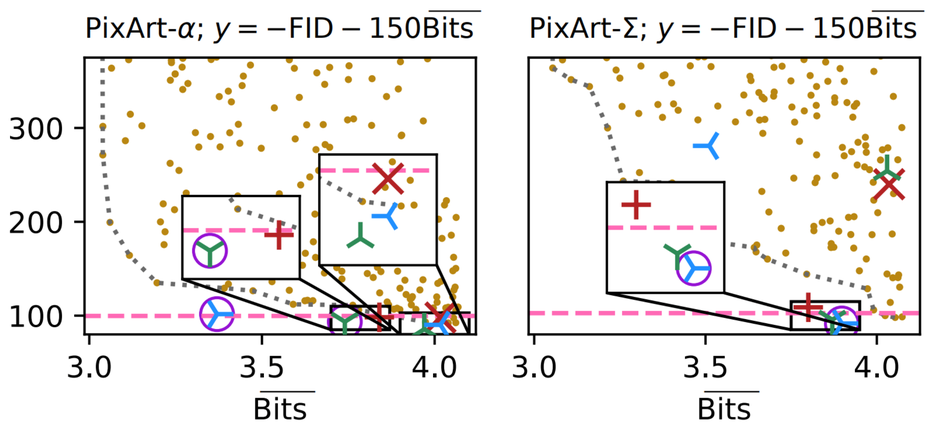
Challenge: How to determine sensitivity?

Solution: Mixed-Precision Search Space

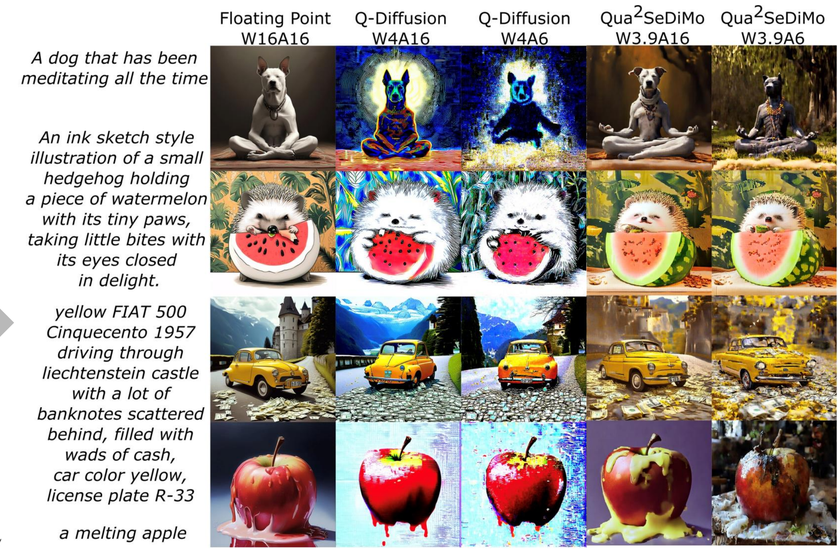


Cast Diffusion Model as search space, where each weight layer (red node) can be quantized to a different bit-precision/method.

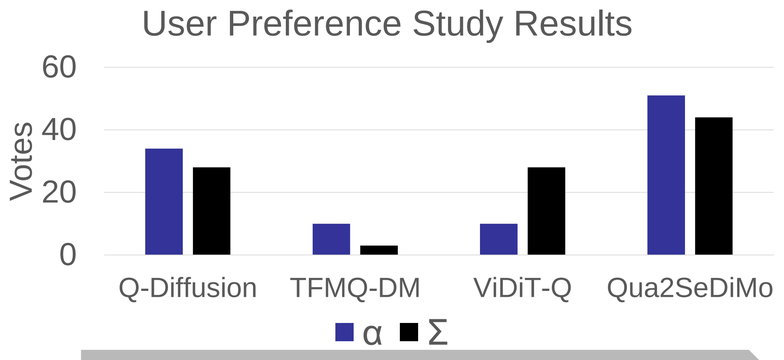
Further, we propose an interpretable, sensitivity insight-extracting and explainable optimization algorithm to *minimize* FID and average bit precision for several models (results in paper).



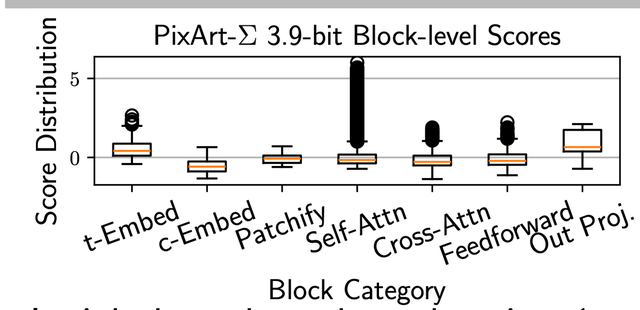
Qualitative & Quantitative Results



We find sub 4-bit quantization configurations that produce images of superior quality.



Extracted Insights



Insight box plots show that time 't-embed' is more sensitive/important than captions 'c-embed'. The self-attention is very important for high-resolution PixArt-Σ, as is the last layer.