

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.



Qua²SeDiMo: Quantifiable Quantization Sensitivity of Diffusion Models

Keith G. Mills¹, Mohammad Salameh², Ruichen Chen¹, Negar Hassanpour², Wei Lu³ and Di Niu¹

¹Dept. ECE, University of Alberta

²Huawei Technologies Canada

³Huawei Kirin Solution

Diffusion Model Post-Training Quantization

Our Hypothesis: Weight Sensitivity

Qualitative & Quantitative Results

Q-Diffusior

W4A6

Q-Diffusior

W4A16

Floating Point

W16A16

An ink sketch style

llustration of a small

hedgehog holding

4.0

Bits

Qua²SeDiMo

N3.9A6

W3.9A16

- Reduce DNN bit-precision. \bullet
 - 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.

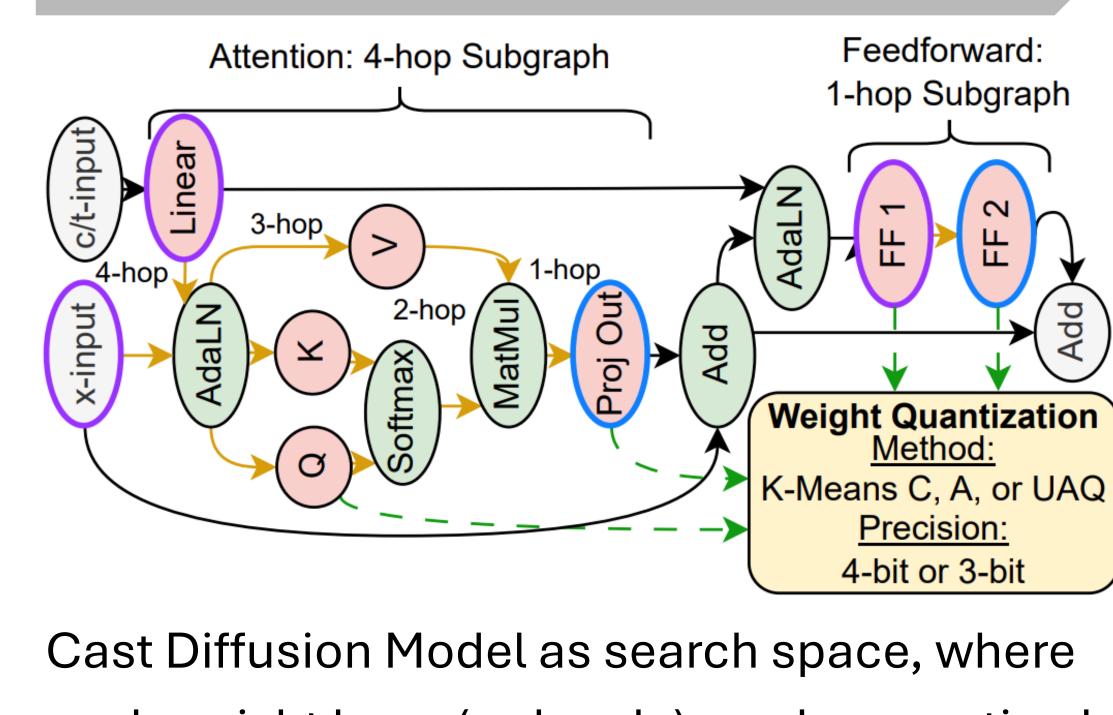
Challenge: <4-bit Weight Quantization

Weight precision controls model size. Most methods achieve 4-bit (W4) weight precision: K-Means **Full Precision** Q-Diffusion UAQ

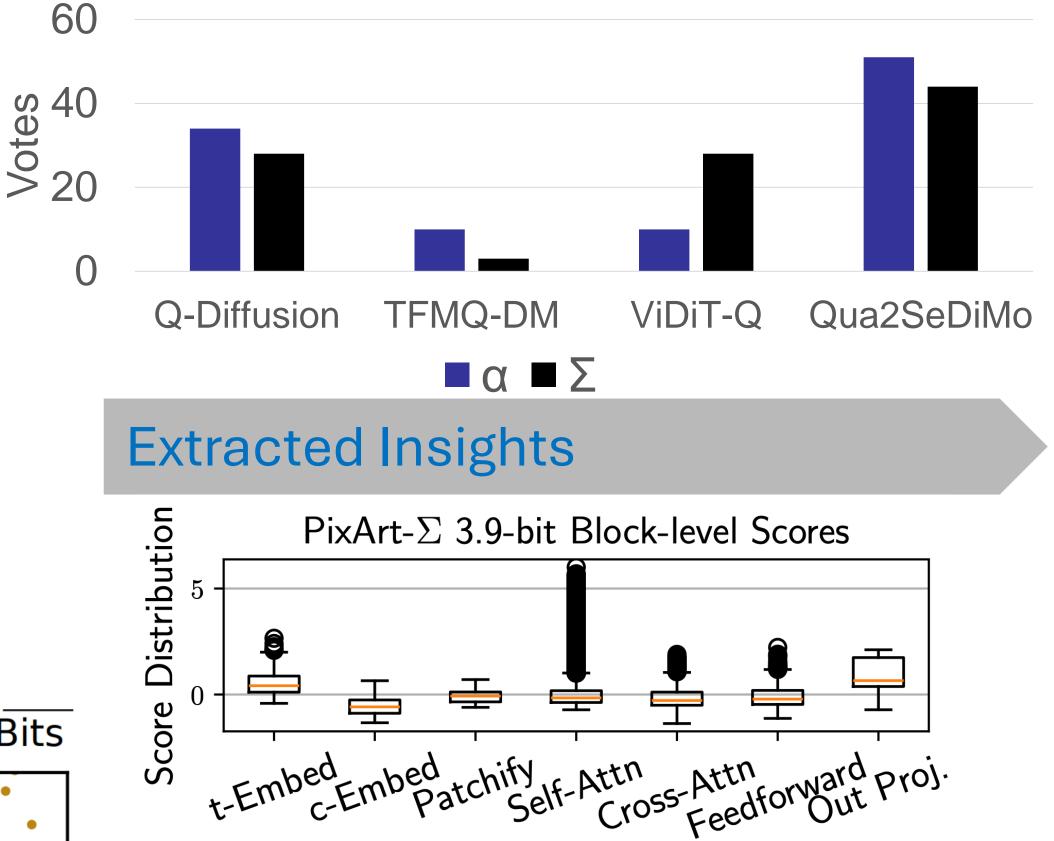


Performance degradation caused by some A dog that has been weight types being quantized to 3-bit meditating all the time precision. Meanwhile, other weights are less crucial and can be lowered to 3-bits. Challenge: How to determine sensitivity?

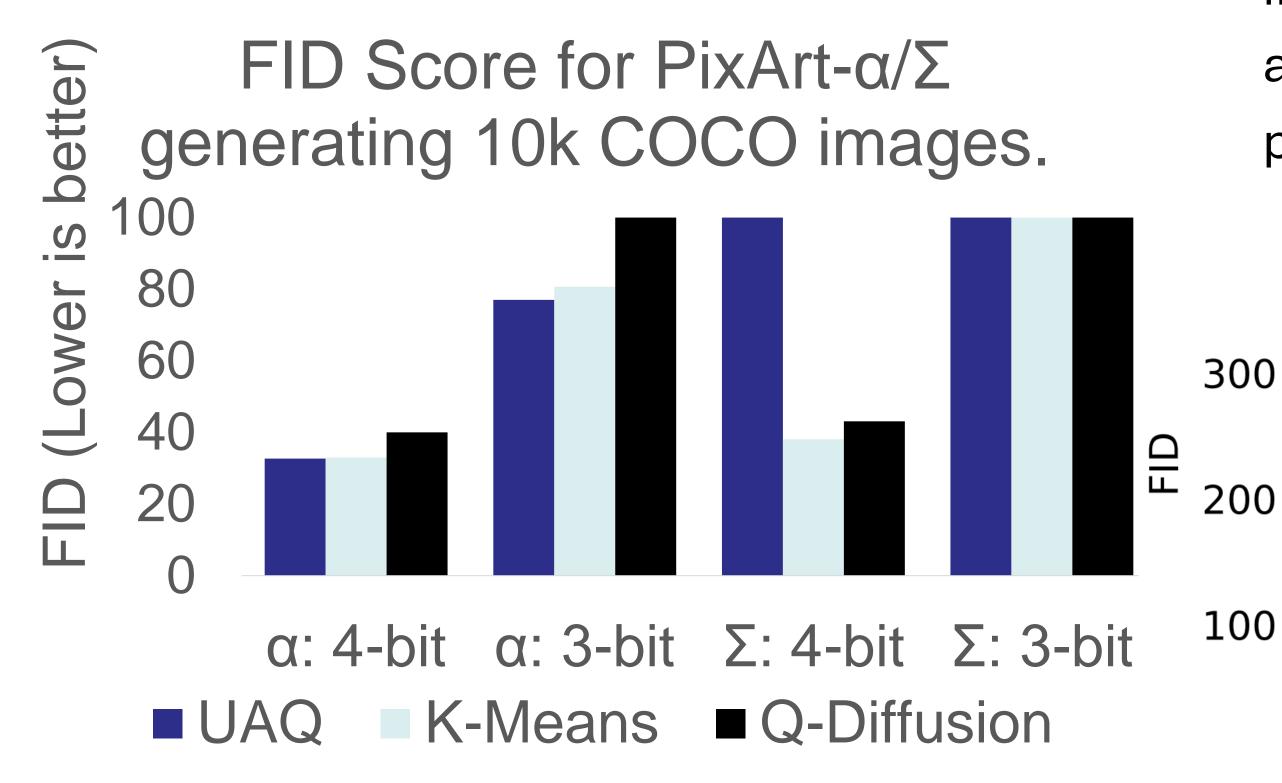
Solution: Mixed-Precision Search Space



a piece of watermelon ith its tiny paws taking little bites with its eyes closed in delight. yellow FIAT 500 Cinquecento 1957 driving through liechtenstein castl with a lot of banknotes scattered 😹 behind, filled with wads of cash, car color yellow, license plate R-33 a melting apple We find sub 4-bit quantization configurations that produce images of superior quality. User Preference Study Results



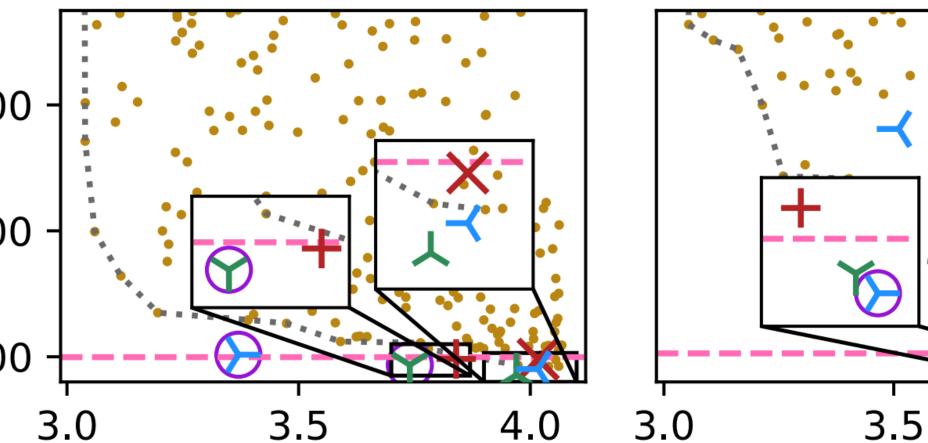
However, 3-bit precision is a struggle.



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).

PixArt- α ; y = -FID - 150BitsPixArt- Σ ; y = -FID - 150Bits



Bits

Block Category Insight box plots show that time 'tembed' is more sensitive/important than captions 'c-embed'. The selfattention is very important for highresolution PixArt- Σ , as is the last layer.