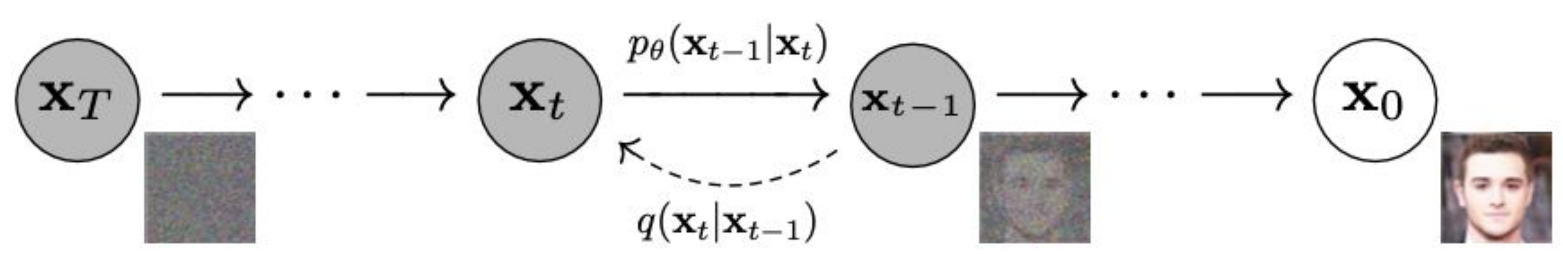
Diffusion Models Quantization

Xiuyu Li

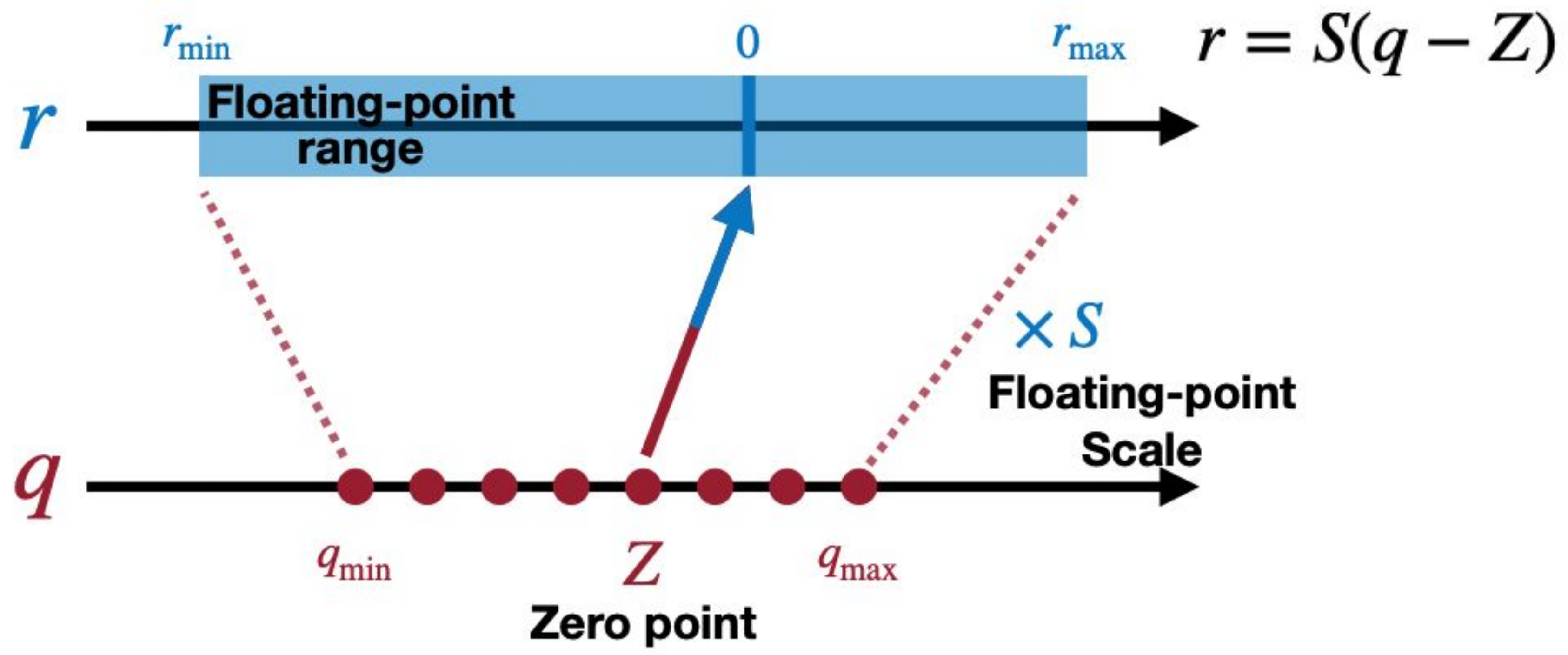
Background: Diffusion Models



- **Diffusion models** slowly add random noise to data and then learn to reverse the diffusion process to construct desired data samples from the noise.
  - core technology behind **AIGC** applications (i.e. text-to-image generation, images inpainting…)
- **Inference is slow** – take several seconds for a single image, while previous SOTA methods (e.g. GANs) generate multiple images under 1s.
- **Memory consumption is high** – stable diffusion has a 860M parameters UNet and takes 7.7 (4.5) GB GPU VRAM to generate an image under FP32 (FP16) precision.

Background: Post-Training Quantization

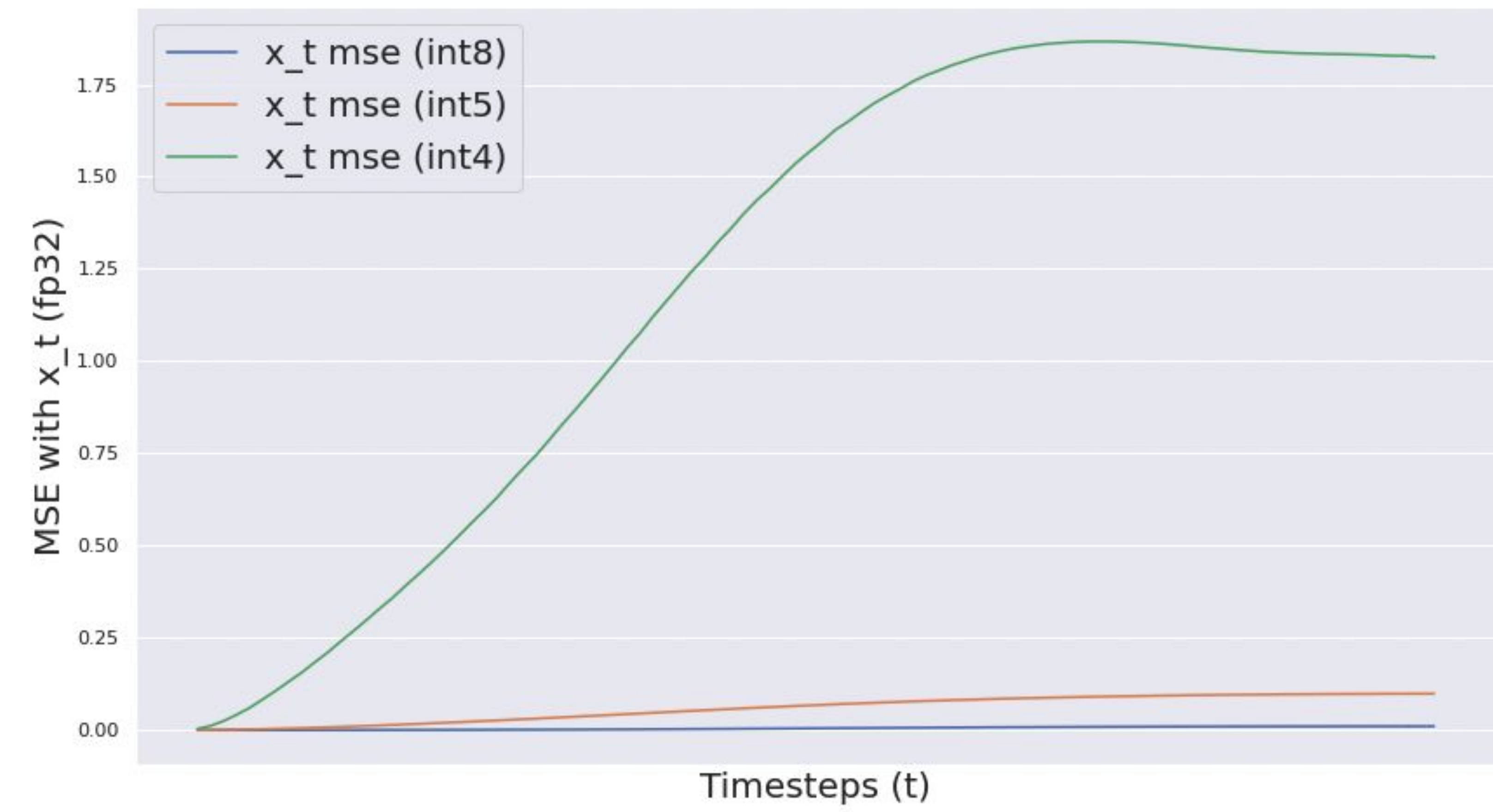
- **Quantization** convert weights and activations to lower bit formats and reduce time and memory consumption



- **Post-training quantization (PTQ)** directly quantizes well-trained models without retraining
  - need training data to calibrate quantized models, usually unavailable due to privacy issues

Quantize diffusion models to reduce memory consumption and accelerate inference speed

- **Property 1:** denoising process has multiple timesteps – feeding model with previous output x_{t-1} at each t > 0, **quantization errors can accumulate**



- **Property 2:** model at **different timestep has different sensitivity** w.r.t quantization
- **Property 3:** we can sample gaussians to generate data with FP32 model for calibration – always **data-free**

Calibrate quantized models with samples from different timesteps

- Naively using SOTA data-free PTQ methods (e.g. SQuant) greatly undermines images quality under **INT4**
- Measure timesteps importance using Peak signal-to-noise ratio (**PSNR**)
- Calibrate mode using a **hessian-based optimization** with weighted sampled data from multiple timesteps

FP32 INT8 INT4 INT4 w/ simple calibration INT4 (ours)

Quantitative results

- Fréchet inception distance (FID) on CIFAR10

FP32	INT8 (Linear)	INT5 (Linear)	INT5 (SQuant)	INT4 (Linear)	INT4 (SQuant)	INT4 (Ours)
5.07	5.93	42.56	28.03	176.88	190.28	13.79

- Model size reduction: scale linearly with #bits e.g. FP16 –> INT4 can usually reduce the size by **2-4x**
- Speed-up: largely dependent on the architecture / weights and activations precisions. But FP16 –> INT4 usually can have **around 2-3x** speedup

Qualitative example: Stable Diffusion

FP32 INT8 INT4

Prompt: *a photograph of an astronaut riding a horse*

Next steps

- Investigate the optimal timesteps importance sampling for the calibration process
- More stable diffusion results
- Use mixed-precision to further lower bits
- Implement customized CUDA kernels to measure the real speed-up in wallclock time

References

[1] Ho et al. Denoising Diffusion Probabilistic Models. 2020
[2] Rombach et al. High-Resolution Image Synthesis with Latent Diffusion Models. 2022
[3] Lee. What are Diffusion Models? 2021.
[4] Li et al. BRECQ: Pushing the Limit of Post-Training Quantization by Block Reconstruction. 2021
[5] Guo et al. SQuant: On-the-Fly Data-Free Quantization via Diagonal Hessian Approximation. 2022