Modulated Diffusion: Accelerating Generative Modeling with Modulated Quantization

Weizhi Gao[1], Zhichao Hou[1], Junqi Yin[2], Feiyi Wang[2], Linyu Peng[3], Xiaorui Liu[1]
[1]North Carolina State University, [2]Oak Ridge National Lab, [3]Keio University

Problem and Challenges

Diffusion Models. Diffusion models consist of a forward process and a backward process, operating over T steps -- **Expensive**.

Forward Process: adding noise for training

Reverse Process: predicting noise for generation

Post-Training Quantization. Quantization reduces the inference cost of models by utilizing low-precision integers. Post-Training Quantization (PTQ) post process models -- **Activation Bottleneck**.

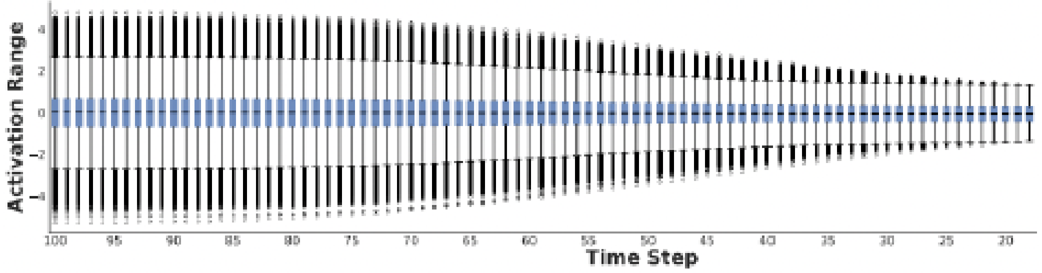
Quantize: converting floating numbers to integers

$$\boldsymbol{x}_{int} = clamp(\lfloor \frac{\boldsymbol{x}}{s} \rceil + \boldsymbol{z}; 0, 2^b - 1)$$

Dequantize: converting integers to floating numbers

$$Q(\boldsymbol{x}) = s(\boldsymbol{x}_{int} - \boldsymbol{z})$$

- The range of activations is *dynamic* in time step and instance.
- The activations is *long-tailed*. [1]

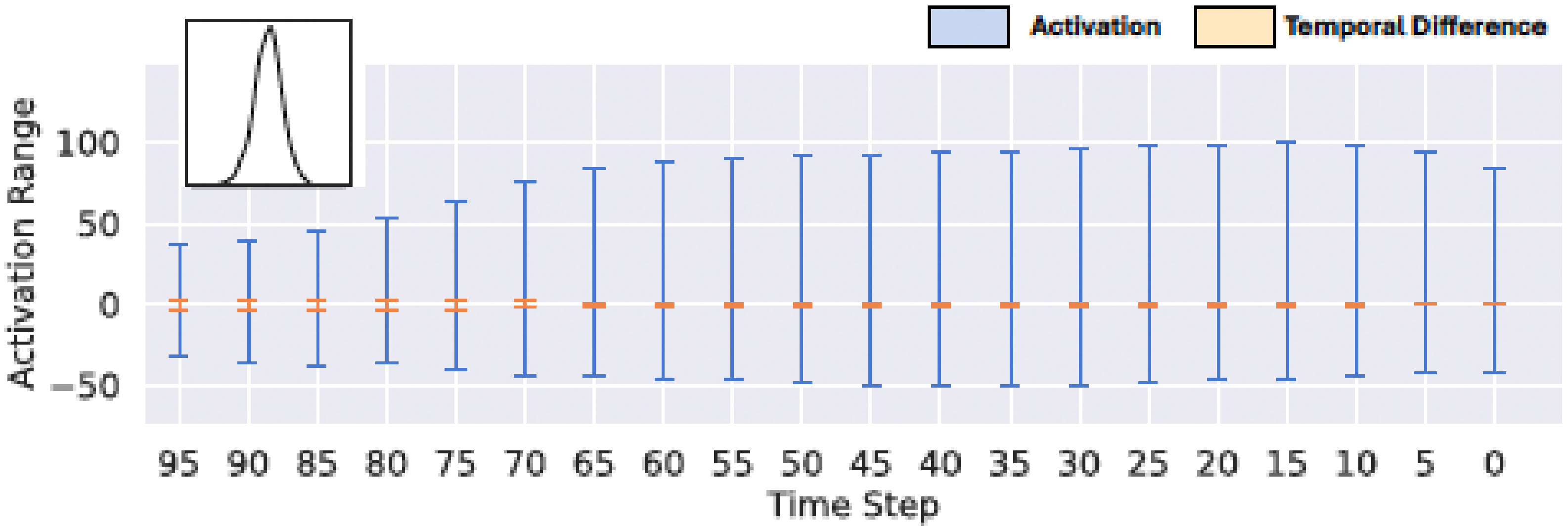


The mismatch between the weights and activations limits the acceleration of quantization on hardware (for instance, W4A8)

[1] Li, Xiuyu, et al. "Q-diffusion: Quantizing diffusion models." *ICCV* 2023.

Preliminary Studies and Motivation

Redundancy across Time Steps. The temporal differences $(\boldsymbol{a}_t^{(l)} - \boldsymbol{a}_{t-1}^{(l)})$ of activations present concentrated and unified distribution compared to the raw activation $(\boldsymbol{a}_t^{(l)})$ at layer l in time step t.



Motivation. Make use of the redundancy among time steps for better quantization acceleration – towards aligned and low bits.

Method: Modulated Quantization

Modulated Quantization. We reformulate the linear operators, such as linear layers and convolutional layers, then we propose a unified framework, MoDiff, that inheriting the advantages of quantization and caching methods, compatible to different solvers.

Reformulation:

$$
\begin{aligned}
\mathbf{o}_T^{(l)} &= \mathcal{A}^{(l)}(\mathbf{a}_T^{(l)}) \\
\mathbf{o}_{T-1}^{(l)} &= \mathcal{A}^{(l)}(\mathbf{a}_{T-1}^{(l)}) = \mathcal{A}^{(l)}(\mathbf{a}_{T-1}^{(l)} - \mathbf{a}_T^{(l)}) + \mathbf{o}_T^{(l)} \\
&\cdots \\
\mathbf{o}_t^{(l)} &= \mathcal{A}^{(l)}(\mathbf{a}_t^{(l)}) \quad = \mathcal{A}^{(l)}(\mathbf{a}_t^{(l)} - \mathbf{a}_{t+1}^{(l)}) + \mathbf{o}_{t+1}^{(l)} \\
&\cdots \\
\mathbf{o}_1^{(l)} &= \mathcal{A}^{(l)}(\mathbf{a}_1^{(l)}) \quad = \mathcal{A}^{(l)}(\mathbf{a}_1^{(l)} - \mathbf{a}_2^{(l)}) + \mathbf{o}_2^{(l)}
\end{aligned}
$$

Since we make use of the linearity of $\mathcal{A}$

$$
\begin{aligned}
\mathcal{A}^{(l)}(\mathbf{a}_t^{(l)}) &= \mathcal{A}^{(l)}(\mathbf{a}_t^{(l)}) - \mathcal{A}^{(l)}(\mathbf{a}_{t+1}^{(l)}) + \mathcal{A}^{(l)}(\mathbf{a}_{t+1}^{(l)}) \\
&= \mathcal{A}^{(l)}(\mathbf{a}_t^{(l)} - \mathbf{a}_{t+1}^{(l)}) + \mathbf{o}_{t+1}^{(l)}.
\end{aligned}
$$

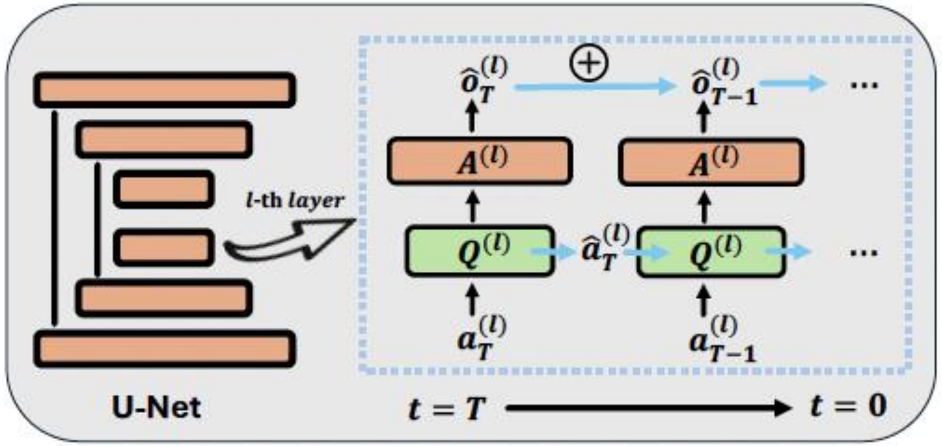
Quantization:

$$
\begin{aligned}
\hat{\mathbf{o}}_T &= \mathcal{A}\Big(Q(\mathbf{a}_T)\Big) && \approx \mathcal{A}(\mathbf{a}_T) \\
\hat{\mathbf{o}}_{T-1} &= \mathcal{A}\Big(Q(\mathbf{a}_{T-1} - \mathbf{a}_T)\Big) + \hat{\mathbf{o}}_T && \approx \mathcal{A}(\mathbf{a}_{T-1}) \\
&\cdots \\
\hat{\mathbf{o}}_t &= \mathcal{A}\Big(Q(\mathbf{a}_t - \mathbf{a}_{t+1})\Big) + \hat{\mathbf{o}}_{t+1} && \approx \mathcal{A}(\mathbf{a}_t) \\
&\cdots \\
\hat{\mathbf{o}}_1 &= \mathcal{A}\Big(Q(\mathbf{a}_1 - \mathbf{a}_2)\Big) + \hat{\mathbf{o}}_2 && \approx \mathcal{A}(\mathbf{a}_1)
\end{aligned}
$$

Method: Error-Compensated Modulation

Implicitly trace quantization error of modulated quantization.

$$
\begin{aligned}
\mathbf{e}_t &= (\mathbf{a}_t - \hat{\mathbf{a}}_{t+1}) - Q(\mathbf{a}_t - \hat{\mathbf{a}}_{t+1}) \\
&= (\mathbf{a}_t - \hat{\mathbf{a}}_{t+1}) - (\hat{\mathbf{a}}_t - \hat{\mathbf{a}}_{t+1}) = \mathbf{a}_t - \hat{\mathbf{a}}_t
\end{aligned}
$$



Error-Compensated Modulation. Trace quantization error to reduce the error accumulation in modulated quantization.

$$
\begin{aligned}
\hat{\mathbf{a}}_T &= Q(\mathbf{a}_T) \\
\hat{\mathbf{o}}_T &= \mathcal{A}(\hat{\mathbf{a}}_T) \\
\hat{\mathbf{a}}_{T-1} &= Q(\mathbf{a}_{T-1} - \hat{\mathbf{a}}_T) + \hat{\mathbf{a}}_T \\
\hat{\mathbf{o}}_{T-1} &= \mathcal{A}(\hat{\mathbf{a}}_{T-1}) = \mathcal{A}\Big(Q(\mathbf{a}_{T-1} - \hat{\mathbf{a}}_T)\Big) + \hat{\mathbf{o}}_T \\
&\dots \\
\hat{\mathbf{a}}_t &= Q(\mathbf{a}_t - \hat{\mathbf{a}}_{t+1}) + \hat{\mathbf{a}}_{t+1} \\
\hat{\mathbf{o}}_t &= \mathcal{A}(\hat{\mathbf{a}}_t) = \mathcal{A}\Big(Q(\mathbf{a}_t - \hat{\mathbf{a}}_{t+1})\Big) + \hat{\mathbf{o}}_{t+1} \\
&\dots \\
\hat{\mathbf{a}}_1 &= Q(\mathbf{a}_1 - \hat{\mathbf{a}}_2) + \hat{\mathbf{a}}_2 \\
\hat{\mathbf{o}}_1 &= \mathcal{A}(\hat{\mathbf{a}}_1) = \mathcal{A}\Big(Q(\mathbf{a}_1 - \hat{\mathbf{a}}_2)\Big) + \hat{\mathbf{o}}_2
\end{aligned}
$$

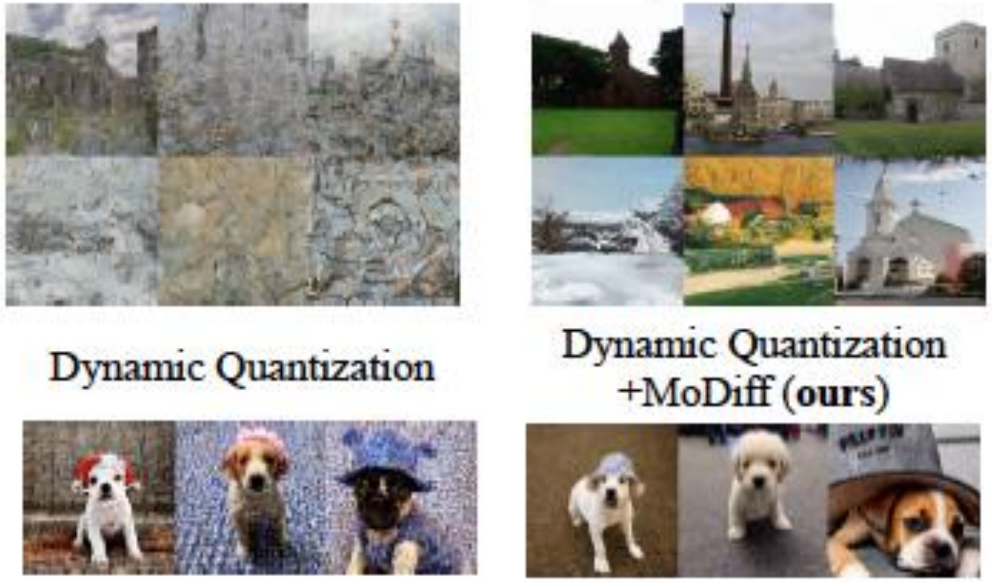
- Theory 1: modulated quantization introduces smaller quantization error by reducing the magnitude of inputs.
- Theory 2: error compensation reduce the accumulated quantization error in an exponential ratio.
- Caching methods are the special cases with 0 bits.
- Our work is **orthogonal** to existing PTQ methods.

Experimental Results and Visualization

LDM-4 on LSUN-Church

Methods	Bits (W/A)	GBops	FID ↓	sFID ↓
Full Prec. (Act)	8/32	5015	4.03	10.89
Q-Diff	8/8	1254	4.24	10.57
Q-Diff+MoDiff (Ours)			**3.85**	10.82
LCQ			4.02	11.53
LCQ+MoDiff (Ours)			3.99	**10.06**
Q-Diff	8/6	1254	55.13	30.98
Q-Diff+MoDiff (Ours)			5.43	13.41
LCQ			4.50	12.90
LCQ+MoDiff (Ours)			**3.89**	**10.12**
Q-Diff	8/4	1254	355.85	187.56
Q-Diff+MoDiff (Ours)			**3.97**	11.16
LCQ			198.37	161.03
LCQ+MoDiff (Ours)			34.02	**10.59**
Q-Diff	8/3	1254	367.51	354.59
Q-Diff+MoDiff (Ours)			**5.40**	**13.81**
LCQ			341.62	407.68
LCQ+MoDiff (Ours)			12.05	35.29

Visualization of LSUN-Church (W4A4)
and MS-COCO (W8A6)



Thanks for your attention!