Q-DiT: Accurate Post-Training Quantization for Diffusion Transformers

Supplementary Material

1. Dynamic Quantization Overhead

The overhead of dynamic quantization is negligible, as we fuse the quantization process into preceding layers, such as LayerNorm. Benchmarks for the *LayerNorm* kernel on a RTX 4090 GPU are presented in Tab.7 Additionally, the INT8 GEMM kernel takes 26.01µs to execute. This indicates that the overhead of dynamic quantization accounts for only 1.3% of the dense GEMM kernel's cost, *demonstrating that dynamic quantization introduces minimal computational overhead.*

Table 7. Measurement of dynamic quantization overhead

Metric	w/o quant	w/ static quant	w/ dynamic quant
Latency↓	5.84µs	6.62µs	6.98µs

2. Additional Observation

Fig. 6 and Fig. 7 illustrate the mean and standard deviation of activations across different labels, highlighting the presence of sample-wise variations in the activation distributions in addition to timestep-wise differences.

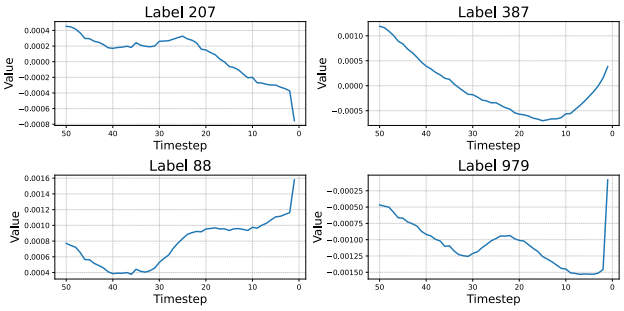

Figure 6. Mean activation values in the blocks.4.attn layer of DiT-XL/2 over 50 timesteps, measured while generating one image per label.

3. Additional Experiment Results

3.1. Visualization of Group Size Configuration

The group size configurations for different models are shown in Fig. 8. We have observed that for different models (e.g., different generation resolutions, different cfg scales), the optimal grouping strategies are quite different, e.g., for higher generation resolution, the top layers have larger group sizes, while for lower generation resolution, the middle layers exhibit the requirements of larger group sizes.

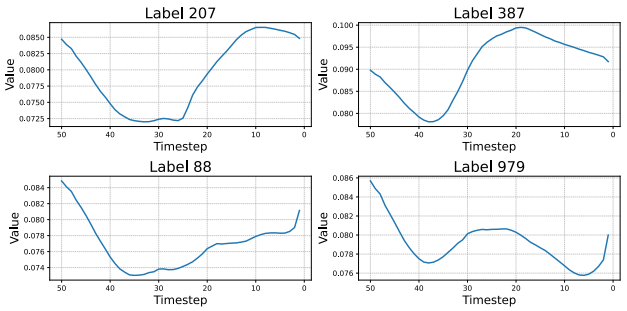

Figure 7. Std of activation values in the blocks.16.mlp layer of DiT-XL/2 over 50 timesteps, measured while generating one image per label.

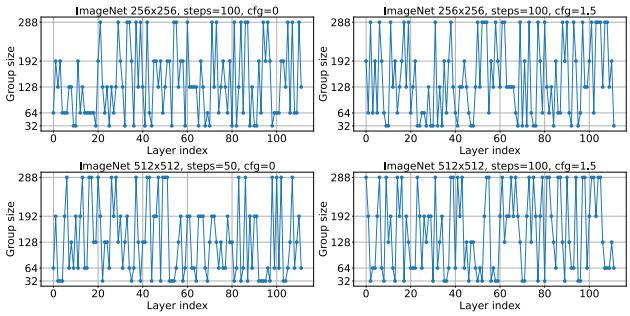

Figure 8. Group size configuration for models at different resolutions, steps and cfg scales.

3.2. Qualitative Results

Image generation results. The additional generated images are presented in Fig. 9. Under the W4A8 setting, the G4W+P4A method produces images of noticeably lower quality, characterized by significant blurring and artifacts. In contrast, our proposed method, Q-DiT, effectively preserves image quality, delivering clear and well-defined visuals even under these challenging conditions.

3.3. Quantitative Results

Additional quantitative results for image generation experiments conducted on ImageNet 512×512 are shown in Tab. 8. Notably, even under the W6A8 setting, methods such as PTQ4DM, RepQ-ViT, and G4W+P4A exhibit significant performance degradation compared to our approach. Furthermore, under the more challenging W4A8 setting, our method continues to maintain high image generation quality, demonstrating its robustness.

(a) G4W+P4A (W4A8) (b) Q-DiT (W4A8)

(c) G4W+P4A (W4A8) (d) Q-DiT (W4A8)

Figure 9. Samples generated by G4W+P4A (one of our baseline) and Q-DiT with W4A8 on ImageNet 256×256 (Top) and ImageNet 512×512 (Bottom).

Table 8. Results on image generation. We show the quantization results of DiT-XL/2 on ImageNet 256×256 and 512×512. 'W/A' indicates the bit-width of weight and activation, respectively.

Model	Bit-width (W/A)	Method	Size (MB)	FID ↓	sFID ↓	IS ↑	Precision ↑
DiT-XL/2 512×512 (steps = 50)	16/16	FP	1349	16.01	20.50	97.79	0.7481
	6/8	PTQ4DM	508	21.22	20.11	80.07	0.7131
		RepQ-ViT	508	19.67	22.35	75.78	0.7082
		TFMQ-DM	508	20.99	22.01	71.08	0.6918
		PTQ4DiT	508	19.42	21.94	77.35	0.7024
		G4W+P4A	520	19.55	22.43	85.56	0.7158
		Ours	517	**16.21**	**20.41**	**96.78**	**0.7478**
	4/8	PTQ4DM	339	131.66	75.79	11.54	0.1847
		RepQ-ViT	339	105.32	65.63	18.01	0.2504
		TFMQ-DM	339	80.70	59.34	29.61	0.2805
		PTQ4DiT	339	35.82	28.92	48.62	0.5864
		G4W+P4A	351	26.58	24.14	70.24	0.6655
		Ours	348	**21.59**	**22.26**	**81.80**	**0.7076**

3.4. Visualization of VBench Results

Fig. 10 presents a radar plot illustrating the VBench results. To facilitate a clear comparison, the values for each dimension have been normalized. Our method, Q-DiT, outperforms the baseline G4W+P4W in 15 out of 16 metrics, closely aligning with the results achieved using full precision.

4. Comparison with Other PTQ Methods

Tab. 9 presents a comprehensive comparison of Q-DiT against various PTQ methods. The W4A8 FID degradation is calculated as the ratio of the FID loss in the quan-

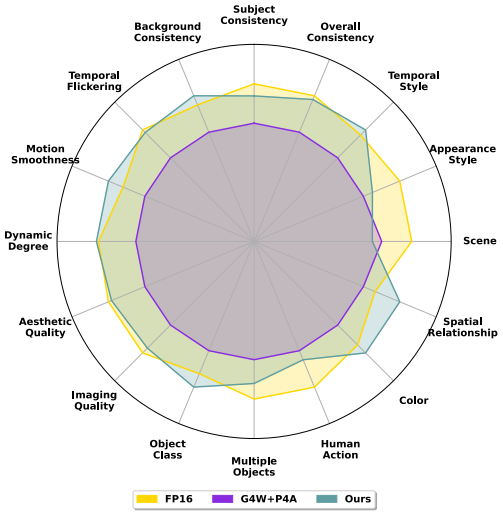

Figure 10. Visualization of Vbench results.

tized results relative to the full precision results on ImageNet 256×256. Notably, PTQ4DM and RepQ-ViT exhibit significant degradation, with FID losses exceeding 1000%, indicating a substantial failure to preserve image quality. PTQ4ViT, a quantization method specifically designed for DiTs, reduces the FID loss to 74.7%. ViDiT-Q, another DiT-focused method, also achieves reduced FID loss but is less practical due to its timestep-wise mixed precision, making it hardware-unfriendly. In contrast, our proposed Q-DiT method not only minimizes FID degradation to just 24.7% but also remains hardware-friendly, offering a significant advantage over other PTQ methods.

Table 9. PTQ method comparisons.

Method	Gradient free	Fine-grained quantization	W4A8 FID degradation	Hardware friendly
PTQ4DM	×	×	1310.7%	✓
RepQ-ViT	✓	×	1170.9%	✓
PTQ4DiT	✓	×	74.7%	✓
ViDiT-Q	✓	×	–	×
Q-DiT (Ours)	✓	✓	24.7%	✓