Memory-Efficient Generative Models via Product Quantization

Supplementary Material

1. Calculation of codebook compression

In Section 3.2 of the main paper, we define the compressed codebook size as $N' = \frac{mn}{16d^2}$ and claim that, after compression, the codebook size is at most one-fourth the size of the quantized weights. To substantiate this claim, we perform calculations using a specific example rather than relying solely on theoretical analysis.

We consider a linear layer from the AdaLN module, with dimensions 6912×1152. After grouping, where 4 numbers are combined into one group, the size is transformed into $6912 \times 288 \times 4$, with each value stored in FP32 (32 bits per number). Following compression, the weights are reduced to $6912 \times 288 \times 1$, where each value is stored as Int8 (8 bits per number). This results in a total of *15,925,248 bits*.

For the compressed codebook, N' is calculated to be *31,104*. The size of the compressed codebook is $31,104 \times 4$, with each value stored in FP16 (16 bits per number), resulting in a total of *1,990,656 bits*. Additionally, the projection matrix from the original codebook to the compressed codebook has dimensions $288 \times 256 \times 1$, with each value stored in Int16 (16 bits per number), amounting to *1,179,648 bits*. In total, the size of the compressed codebook and the projection matrix is less than one-fourth of the size of the compressed weights, thereby validating our claim.

2. Experiments

2.1. Experiment details

Class conditional generation. The original DiT-XL/2 model checkpoint is provided in the DiT repository. The quantized models are evaluated on the ImageNet validation set using ADM's evaluation suite. Specifically, we use pre-computed sample batches from ImageNet as a reference. These reference batches include pre-computed statistics over the entire dataset and 10,000 images for calculating precision and recall.

For compression stage, we adopt a stochastic relaxation of k-means with decreasing noise during the process. During fine-tuning, we set the weight decay to 0 and the class dropout probability to 0.1, which means the image labels are randomly dropped during training for better unconditional generation and classifier-free guidance.

Our code is built upon the repositories of DiT [5] and PQF [4], where DiT provides the evaluation and training framework, and PQF offers a basic implementation of VQ. The PQ implementation references the approach for CNNs from [7]. Calibration is conducted using the ImageNet dataset [1]. We sincerely appreciate the contributions of the open-source community. The LDM code for both training and validation is based on the LDM repository [6], while the PQ for convolution follows the VQ for CNN implementation from [4].

Text-to-Image Generation. The T2I model code is built on diffusers [8], with model checkpoints sourced from Hugging Face. The evaluation toolkit is developed based on GenEval [2] and T2I-CompBench [3].

2.2. More experiments and visualization

We present experimental results for a resolution of 512×512. Our compression method is implemented and compared against other approaches. For evaluation, we use a DDPM scheduler with 50, 100, and 250 timesteps, along with a default classifier-free guidance (CFG) scale of 1.5. The results are summarized in Table 1, where our method consistently outperforms others, demonstrating significant improvements across a wide range of metrics.

Additionally, we provide visualizations of generated images under 1-bit and 2-bit compression with varying timesteps. These results are illustrated in Figure 1.

Table 1. Performance comparison on ImageNet with resolution of 512×512

Timesteps	Bit-width	Method	Size ratio	FID ↓	sFID ↓	IS ↑	Precision ↑
100	32	FP	-	5.00	19.02	274.78	0.8149
	2	GPTQ	10.10 ×	3.1e2	1.7e2	2.66	0.0179
	2	Q-DiT	10.23 ×	3.8e2	2.2e2	1.25	0.0001
	2	VQ4DiT	10.59 ×	34.32	51.08	57.03	0.7929
	2	Ours	9.92 ×	19.18	26.42	107.14	0.7636
50	32	FP	-	6.02	21.77	246.24	0.7812
	2	GPTQ	10.10 ×	3.2e2	1.8e2	2.65	0.0170
	2	Q-DiT	10.23 ×	3.8e2	2.2e2	1.24	0.0001
	2	VQ4DiT	10.59 ×	35.08	48.81	56.82	0.7744
	2	Ours	9.92 ×	22.23	29.71	92.96	0.7694

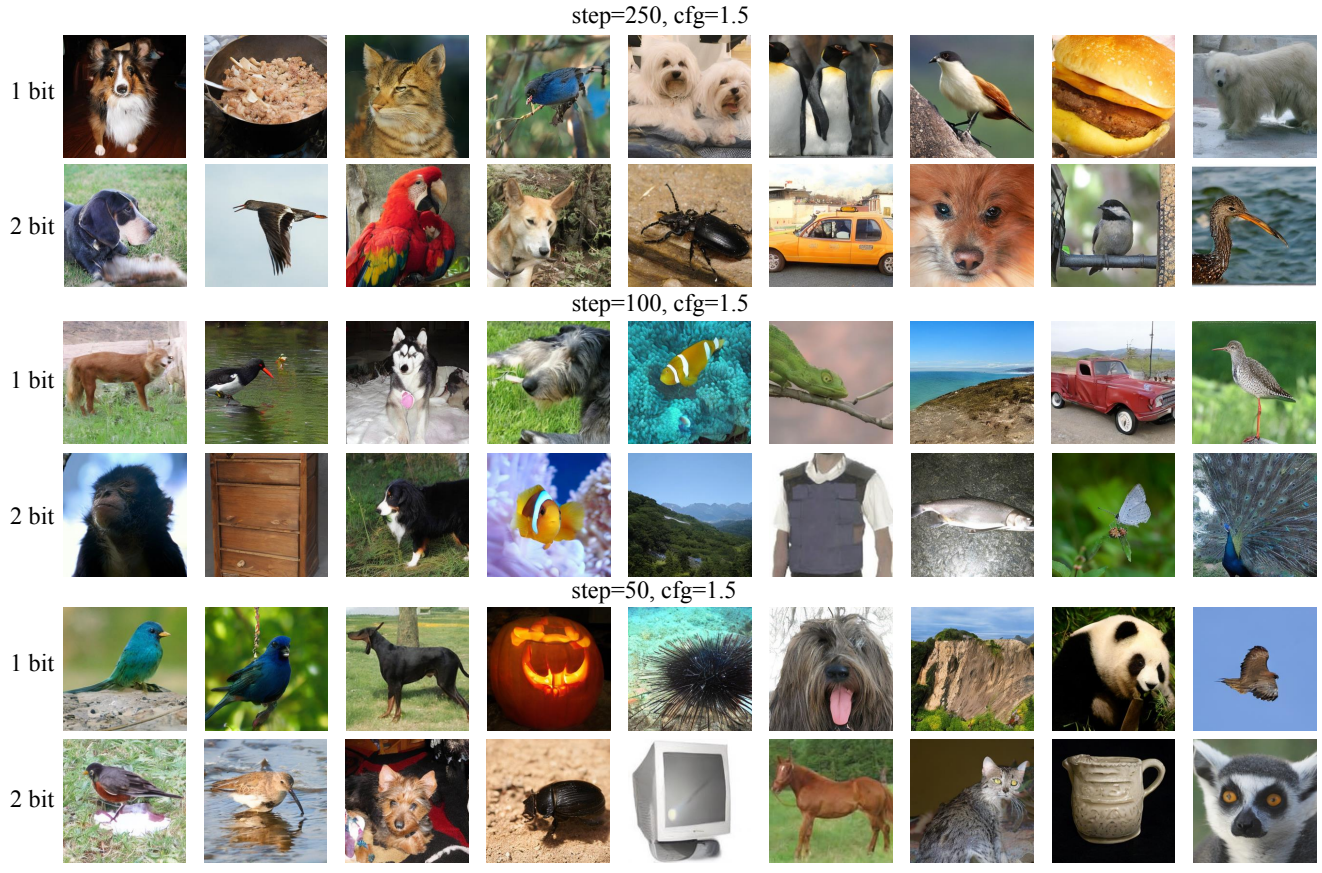

Figure 1. Visualization of generation results from the DPQ-compressed model.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[2] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152, 2023. 1

[3] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in*

Neural Information Processing Systems, 36:78723–78747, 2023. 1

[4] Julieta Martinez, Jashan Shewakramani, Ting Wei Liu, Ioan Andrei Bârsan, Wenyuan Zeng, and Raquel Urtasun. Permute, quantize, and fine-tune: Efficient compression of neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15699–15708, 2021. 1

[5] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 1

[6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[7] Pierre Stock, Armand Joulin, Rémi Gribonval, Benjamin Graham, and Hervé Jégou. And the bit goes down: Revisiting the quantization of neural networks. In *International Conference on Learning Representations*, 2020. 1

[8] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 1