# A. Appendix

## A.1. Implementation Details

For all experiments, unless stated otherwise, we use the following hyperparameters: for classification tasks, 1024 images are generated, and for object detection, 32 images are used. Each result represents an average of five experiments. We perform 1000 optimization iterations, using the RAdam optimizer [36] with an initial learning rate of 16 and a reduced learning rate on the plateau scheduler.

For $\phi_{prep}$, we begin by applying a $3 \times 3$ Gaussian smoothing filter. Then, we randomly apply horizontal flipping to the images with a probability of 0.5. Next, we perform random cropping by starting with images that are 32 pixels larger in both height and width than the final output size and then cropping them to the desired output shape. The final images are center-cropped.

## A.2. Derivation of Equation (5)

Here, we provide the derivation of Eq. (5). We begin with the data mean decomposition, which is followed by the decomposition of variance.

$$\begin{aligned}
\left[\bar{\boldsymbol{\mu}}_l(\boldsymbol{\mathcal{D}})\right]_i &\triangleq \frac{1}{M}\sum_{m=1}^{M}\frac{1}{d_l}\sum_{j=1}^{d_l}\left[\boldsymbol{y}_l^{(m)}\right]_{i,j} \\
&= \frac{1}{N}\sum_{n=1}^{N}\frac{1}{K}\sum_{k\in\mathcal{S}_n}\frac{1}{d_l}\sum_{j=1}^{d_l}\left[\boldsymbol{y}_l^{(k)}\right]_{i,j} \\
&= \frac{1}{N}\sum_{n=1}^{N}\left[\bar{\boldsymbol{\mu}}_l(\boldsymbol{B}_n)\right]_i, \qquad (9)
\end{aligned}$$

where $\mathcal{S}_n$ is the set of index correspond to the $n^{th}$ batch. The first step follows directly from the definition in (3a). In the second step, we apply the linearity of summation. Finally, we use the notion of an empirical mean over batch. Next, we present the relationship between the variance and the second moment:

$$\begin{aligned}
\left[\bar{\boldsymbol{\sigma}}_l(\boldsymbol{\mathcal{D}})\right]_i &\triangleq \frac{1}{M}\sum_{m=1}^{M}\frac{1}{d_l}\sum_{j=1}^{d_l}\left(\left[\boldsymbol{y}_l^{(m)}\right]_{i,j} - \left[\bar{\boldsymbol{\mu}}_l(\boldsymbol{\mathcal{D}})\right]_i\right)^2 \\
&= \left[\bar{\boldsymbol{\mu}}_l(\boldsymbol{\mathcal{D}})\right]_i^2 + \frac{1}{M}\sum_{m=1}^{M}\frac{1}{d_l}\sum_{j=1}^{d_l}\left[\boldsymbol{y}_l^{(m)}\right]_{i,j}^2 - 2\left[\boldsymbol{y}_l^{(m)}\right]_{i,j}\left[\bar{\boldsymbol{\mu}}_l(\boldsymbol{\mathcal{D}})\right]_i \\
&= \frac{1}{N}\sum_{n=1}^{N}\frac{1}{K}\sum_{k\in\mathcal{S}_n}\frac{1}{d_l}\sum_{j=1}^{d_l}\left[\boldsymbol{y}_l^{(m)}\right]_{i,j}^2 - \left[\bar{\boldsymbol{\mu}}_l(\boldsymbol{\mathcal{D}})\right]_i^2 \\
&= \frac{1}{N}\sum_{n=1}^{N}\left[\bar{\boldsymbol{M}}_\ell^{(2)}(\boldsymbol{B}_n)\right]_i - \left[\bar{\boldsymbol{\mu}}_l(\boldsymbol{\mathcal{D}})\right]_i^2, \qquad (10)
\end{aligned}$$

The first step follows from the definition in (3b). In the second step, we expand the parentheses. In the third step, we apply the linearity of summation and the notion of empirical mean. Finally, we use the notion of an empirical moment over the batch.

Table 5. Top-1 accuracy on ImageNet-1k validation set with models quantized using BRECQ [30] in an academic quantization setting. The models are quantized using four different data generation algorithms and real data

| Method | ResNet-18 71.06 | ResNet-50 77.0 | MBV2 72.49 |
|---|---|---|---|
| | W4A4 Academic Quantized | | |
| ZeroQ [4] | 68.67±0.09 | 74.26±0.12 | **69.27**±0.22 |
| IntraQ [60] | 67.78±0.27 | 66.08±0.67 | 69.06±0.13 |
| GENIE [25] | **69.67**±0.05 | **74.90**±0.06 | 69.13±0.05 |
| DGH (Ours) | **69.56**±0.04 | **74.72**±0.12 | **69.23**±0.08 |
| Real Data | 69.74±0.07 | 74.90±0.06 | 69.30±0.10 |
| | W2A4 Academic Quantized | | |
| ZeroQ [4] | 59.73±0.17 | 63.53±0.22 | 27.42±1.04 |
| IntraQ [60] | 49.17±0.56 | 44.03±1.76 | 22.8±3.17 |
| GENIE [25] | **64.37**±0.17 | **69.53**±0.04 | **48.23**±3.24 |
| DGH (Ours) | **64.15**±0.10 | **68.91**±0.06 | **45.83**±1.44 |
| Real Data | 65.86±0.10 | 70.28±0.03 | 54.46±1.44 |

## A.3. Additional Results

### A.3.1 Academic Quantization Results

We validate our approach within the academic quantization scheme. In this setup, the first and last layers and the input to the second layer are quantized to 8 bits, while the output layer remains unquantized. Tab. 5 shows the Top-1 accuracy on the ImageNet-1k validation set using the BRECQ quantization algorithm under the academic quantization scheme, with activation bit-width set to 4 bits and weight bit-width set to 2 and 4 bits. The results demonstrate that the DGH achieves competitive results to GENIE [25], while greatly improving the image generation runtime since DGH does not use a generator. Specifically, GENIE [25] requires approximately two and a half hours on a V100 GPU to generate 1024 images for ResNet18, while DGH takes less than half an hour on an RTX 3090. Although ZeroQ [4] and IntraQ [60] may offer faster generation times than DGH, they deliver inferior performance in both academic and hardware-friendly quantization schemes.

### A.3.2 The Effect of Varying Bit Widths

Here, we provide results offering further insights into the performance of DGH. We present the Top-1 accuracy on ImageNet-1k validation set using the BRECQ quantization algorithm in the hardware-friendly quantization setting, tested across various bit-widths. In this experiment, we keep the weight bit-width fixed at 8 bits and evaluate

IntraQ [60], GENIE [25], and DGH at different activation bit-widths. The results are shown in Fig. 6. Next, we fix the activation bit-width at 8 bits and evaluate IntraQ [60], GENIE [25], and DGH at varying weight bit-widths, with the results presented in Fig. 7. From both sets of results, we observe that the performance gap between DGH and IntraQ [60], GENIE [25] increases as the bit-width decreases. Note that we omit ZeroQ [4] as its accuracy is consistently near zero in all cases.

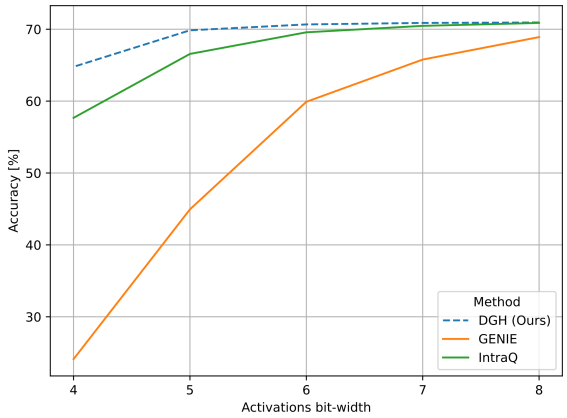


Figure 6. Top-1 accuracy on the ImageNet-1k validation set using the BRECQ quantization algorithm in a hardware-friendly quantization setting, with various activation bit-widths while the weight bit-width is fixed at 8 bits. The y-axis represents the accuracy, and the x-axis represents the activation bit-width.

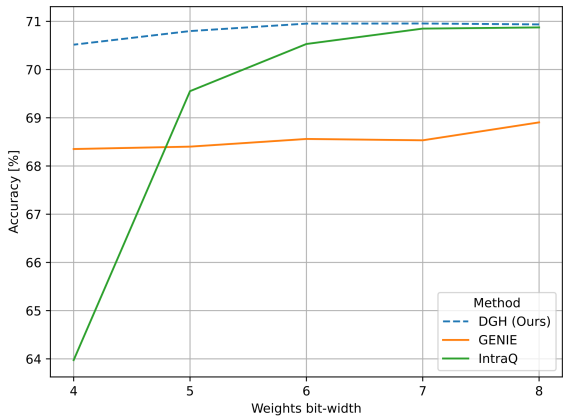


Figure 7. Top-1 accuracy on the ImageNet-1k validation set using the BRECQ quantization algorithm in a hardware-friendly quantization setting, with various weight bit-widths while the activation bit-width is fixed at 8 bits. The y-axis represents the accuracy, and the x-axis represents the weights bit-width.

### A.3.3 Generated Images Using DGH

Next, we visualized the images generated by DGH from ResNet18 after 1k iteration, as shown in (Fig. 8), which were used in all experiments Additionally, we present images generated after 40k iterations in (Fig. 9). In Fig. 8, we observe the initial emergence of shapes and patterns in the generated images. As shown in Fig. 9, with additional iterations, these patterns and class-specific features become increasingly distinct, reflecting improved image quality and clearer representation of different classes in the generated images.

Figure 8. Images generated from a ResNet18 using DGH with 1k Iterations.

Figure 9. Images generated from a ResNet18 using DGH with 40k Iterations.