We thank the reviewers for the positive comments, and would like to provide answers to the questions raised.

We would like to first address Reviewer #1's questions regarding the inference time and a case study for IoT devices. The reviewer states that not all compression techniques provide good computation gains due to a lot of constraints. This is why in our evaluation, the target is a custom hardware implementation, and thus the issues above of efficiently targeting a particular CPU or GPU don't apply on custom hardware. Instead, the #Gates estimations measure hardware efficiency and they are derived from a hardware accelerator design mapped to a Stratix 10 FPGA device. We additionally ensure that all designs with different quantization methods have the same processing performance (line 246 in the paper). A hardware design using focused quantization (FQ) requires fewer gates than its counterparts (LQ-Net, ABC-Net) to achieve the same performance, hence is more power efficient. Conversely, assuming that they all use the same amount of logic resources, FQ thus has a higher throughput than the compared methods, as we can perform more computations in parallel. While it is possible to provide certain performance statistics (*e.g.* energy consumed per image, frames per second given the same chip area, *etc.*) for a particular ASIC design, yet the intricacy involved (*e.g.* the technology node, circuit design, memory bandwidth, *etc.*) is beyond the scope of this paper. The reviewer suggests us to provide an example of how our technique relates to future deployment of CNN accelerators in IoT devices. We believe future IoT systems bring heterogeneous compute: CNN accelerators as co-processors to assist general purpose CPUs. Note that IoT CNN accelerators today typically operate on common quantizations with higher bit-widths, *e.g.* 8-bit fixed-point weights and activations in Eyeriss v2 [1]. The updated paper will address this and include in Table 4 additional #Gates comparison with 8-bit fixed-point weights and activations.

Reviewer #3 suggested that as focused quantization introduces high-precision hyperparameters μ_-, μ_+, σ_- and σ_+, the dot products would require more high-precision multiply-adds. We would like to thank the reviewer for catching this, and will clarify this accordingly. In all of our experiments in the supplementary source code, to simplify inference computation, the hyperparameters μ_- and μ_+ are quantized to the nearest powers-of-two values, and σ_- and σ_+ are constrained to be equal in the optimization process. Effectively in our implementation of the dot-product, all high-precision parameters, *i.e.* σ_-, σ_+ and the layer-wise learnable scalar α, can thus be fused into batch normalization during inference. The reviewer also pointed out that Figure 4 is missing σ values, which can be explained by the same reason above. We will clarify this in both Figure 2 and Figure 4.

In Table 1 below, we present the top-1 and top-5 accuracies for the pruned, encoded but not quantized models, as kindly suggested by Reviewer #3. These results will later be added to the paper. We are working on 3-bit results as kindly requested by Reviewer #2, and will include them in later changes.

Finally, we would like to additionally make the following minor changes to further improve the quality of the paper:

1. The notation of the form "$/N$" (*e.g.* "$/128$") in Figure 2 and Figure 4 denotes all numbers in the same grid share a common denominator N, and thus the true values are the numbers scaled by $\frac{1}{N}$. We will add descriptions to the figures to make it clear to the readers.
2. We will add the y-axis labels ("frequency") to Figure 2 and Figure 3 as kindly suggested by Reviewer #3.
3. We will adopt the numerical citation style, replace "DNNs" in the title with "CNNs", and fix caption spacing and the notation as kindly suggested by Reviewer #1.

We thank the reviewers again for your detailed comments and help on improving the quality of our paper.

Table 1: Additional results for Table 1, newly added rows are shaded.

Model	**Top-1**	**Δ**	**Top-5**	**Δ**	**Sparsity (%)**	**Size (MB)**	**CR (×)**
ResNet-18	68.94	—	88.67	—	0.00	46.76	—
Pruned	69.24	0.30	89.05	0.38	74.86	8.31	5.69
ResNet-50	75.58	—	92.83	—	0.00	93.82	—
Pruned	75.10	-0.48	92.58	-0.25	82.70	11.76	7.98
MobileNet-V1	70.77	—	89.48	—	0.00	16.84	—
Pruned	70.03	-0.74	89.13	-0.35	33.80	6.89	2.44
MobileNet-V2	71.65	—	90.44	—	0.00	13.88	—
Pruned	71.24	-0.41	90.31	-0.13	31.74	5.64	2.46

References

[1] Y. Chen, J. S. Emer, and V. Sze. Eyeriss v2: A flexible and high-performance accelerator for emerging deep neural networks. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 2018.