PTQ4SAM: Post-Training Quantization for Segment Anything

Supplementary Material

This supplementary document is organized as follows: 1) section A: more details on Bimodal Integration (BIG); 2) section B: more quantitative studies of Adaptive Granularity Quantization (AGQ); 3) section C: more qualitative results for instance segmentation.

A. More Details on BIG

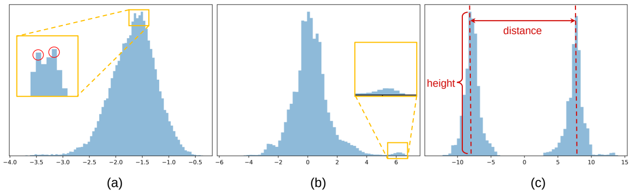

Figure S1. Three typical examples in BIG strategy.

A.1. Bimodal Discovery

As we mentioned in the main paper, we utilize the continuous probability density function to characterize the peaks. However, merely using the naive local maxima will induce an over-detection issue. We summarize the issue in two situations: 1) Two neighboring bumps in one peak are recognized as two peaks (Figure S1(a)). 2) Wrongly consider the small bump as a peak (Figure S1(b)). To address it, we impose constraints stipulating that both the peak height and the distances between two peaks must exceed a predetermined threshold in Figure S1(c). Smaller peaks are removed first until the condition is fulfilled for all remaining peaks.

A.2. Effect of Sign Operation

To verify the effectiveness of our BIG strategy, we show the representative real distributions of query and key activations before and after sign operation. As shown in Figure S2, after sign operation, the bimodal `post-Key-Linear` distribution will be transferred to a normal distribution, narrowing the range from -13∼14 to 3∼14 (row 1). Meanwhile, the query activations remain normal distribution invariantly, slightly reducing the range from -843∼848 to -848∼296 (row 2). Intuitively, our BIG is beneficial for quantization and the sign operation can be performed in advance.

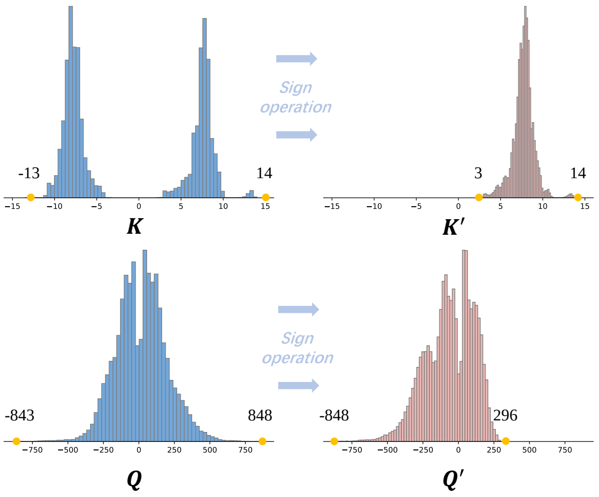

Figure S2. The distribution of query and key activations before and after BIG strategy.

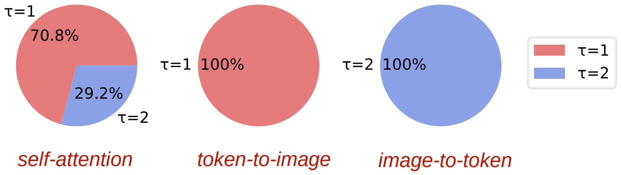

Figure S3. Pie charts depicting the optimal τ across various attention mechanisms in SAM-L.

B. Quantitative Studies of AGQ

We complete the discussion related to the suitable granularity (optimal τ) for different scenarios. As mentioned in Section 3.3, a smaller τ can better quantize lower attention scores. Conversely, with an increment in τ, the higher attention scores can be quantized in a more fine-grained fashion. For simplicity, we conduct a statistical analysis of optimal τ across diverse post-Softmax distributions at W4A4. As illustrated in Figure S3, in `token-to-image`, our AGQ uniformly favors τ=1 because there are more low attention scores (see Figure 1 in the main paper). In `image-to-token`, τ=2 is prominently selected to accurately quantize more high scores. And in `self-attention`, there is a coexistence of τ=1 and τ=2 for the combination of both high and low attention scores.

Model	**SAM-B**		**SAM-L**		**SAM-H**	
#bits	W6A6	W4A4	W6A6	W4A4	W6A6	W4A4
MSE^s	30.2	14.4	35.7	28.3	36.5	32.6
MSE^o	**30.3**	**16.0**	**35.8**	**28.7**	**36.5**	**33.5**

Table S1. Objective test for instance segmentation. ^{s} represents quantization error for post-Softmax activations and ^{o} means quantization error for output activations of matrix multiplication.

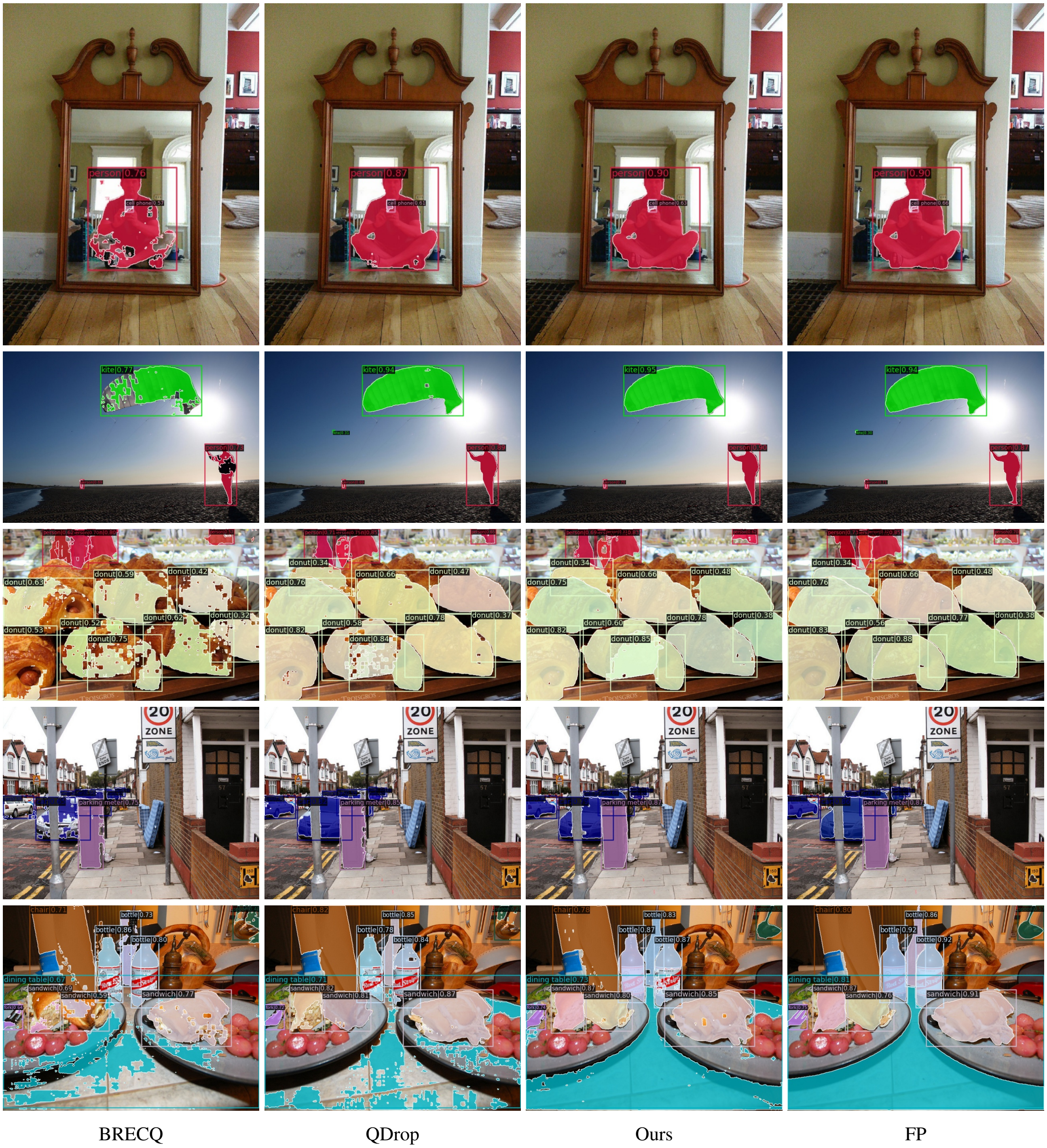

Figure S4. Visualization of instance segmentation on 4-bit SAM-L.

Therefore, our AGQ adopts suitable granularity solutions towards the post-Softmax distribution across diverse attention mechanisms. Additionally, we compare the loss function in Eq. 14 (row 2) with local quantization errors of the attention map $\boldsymbol{A}$ (row 1). Table S1 indicates that Eq. 14 addresses the inconsistent issue and achieves stable performance, especially at low-bit.

C. More Qualitative Results

More instance segmentation results are given in Figure S4 produced by 4-bit BRECQ [1], QDrop [2], PTQ4SAM and full-precision SAM-L. Notably, our model demonstrates superior performance in terms of both completeness and clarity when compared to other methodologies. In a simple scenario with a single object, such as the `person` in row 1 and the `kite` in row 2, our method is capable of providing a more comprehensive description of the object boundaries, without missing any pixels. In cases where objects overlap, as observed in rows 3 and 4, our quantized model accurately distinguishes each individual object and successfully separates them from complex backgrounds. Conversely, other methods often struggle to segment occluded objects accurately, capturing unnecessary details. Particularly when recognizing background objects like the `dining table`, as depicted in row 5, the results obtained from alternative approaches exhibit notable incompleteness. Conversely, our approach excels in effectively identifying the entire object, showcasing a significant advantage over other methods.

References

[1] Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Hu, Qi Zhang, Fengwei Yu, Wei Wang, and Shi Gu. Brecq: Pushing the limit of post-training quantization by block reconstruction. *arXiv preprint arXiv:2102.05426*, 2021. 3

[2] Xiuying Wei, Ruihao Gong, Yuhang Li, Xianglong Liu, and Fengwei Yu. Qdrop: Randomly dropping quantization for extremely low-bit post-training quantization. *arXiv preprint arXiv:2203.05740*, 2022. 3