PQ-SAM: Post-training Quantization for Segment Anything Model Supplementary Material

Xiaoyu Liu[1], Xin Ding[1], Lei Yu[2], Yuanyuan Xi[2], Wei Li[2], Zhijun Tu[2], Jie Hu[2], Hanting Chen[2], Baoqun Yin[1], and Zhiwei Xiong[1(✉)]

[1] University of Science and Technology of China
{liuxyu,xinding64}@mail.ustc.edu.cn, zwxiong@ustc.edu.cn
[2] Huawei Noah's Ark Lab
{wei.lee,zhijun.tu}@huawei.com

1 Results of the 8-bit quantization setting

Table 1: Quantitative Results on the W8A8 Quantizaiton Setting for the point prompt mode of SAM.

	Method	Bit	BBBC038V1		NDD20		Cityscape		DOORS		iShape		NDISPark	
			IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice
ViT-H	FP	-	0.7795	0.8500	0.8468	0.9100	0.5945	0.7061	0.8496	0.9126	0.3722	0.5108	0.8192	0.8938
	Ours	8	0.7885	0.8605	0.8383	0.9044	0.5851	0.7011	0.8396	0.9051	0.3655	0.5077	0.8188	0.8943
ViT-L	FP	-	0.7761	0.8476	0.8468	0.9075	0.5939	0.7068	0.8514	0.9138	0.3479	0.4856	0.8127	0.8893
	Ours	8	0.7647	0.8365	0.8377	0.9034	0.5870	0.7012	0.8452	0.9091	0.3386	0.4780	0.8097	0.8873
ViT-B	FP	-	0.7666	0.8417	0.8322	0.9000	0.5692	0.6884	0.8294	0.8987	0.3536	0.4987	0.7933	0.8760
	Ours	8	0.7368	0.8100	0.8211	0.8924	0.5645	0.6835	0.8374	0.9056	0.3466	0.4912	0.7950	0.8775

In Table 1, we present additional quantitative results for our PQ-SAM method using the 8-bit quantization setting. The results demonstrate that PQ-SAM achieves near-lossless quantization on the majority of the datasets, thereby effectively preserving the ability of SAM.

2 More visualization

In Figure 1 and Figure 2, we provide additional visual results for both the point prompt mode and automatic mode of our proposed method. These results demonstrate the superiority of our PQ-SAM method compared to existing PTQ methods in terms of segmentation performance.

X. Liu and X. Ding—Equal contribution. This research was conducted while Xiaoyu Liu and Xin Ding were interns at Huawei Noah's Ark Lab.

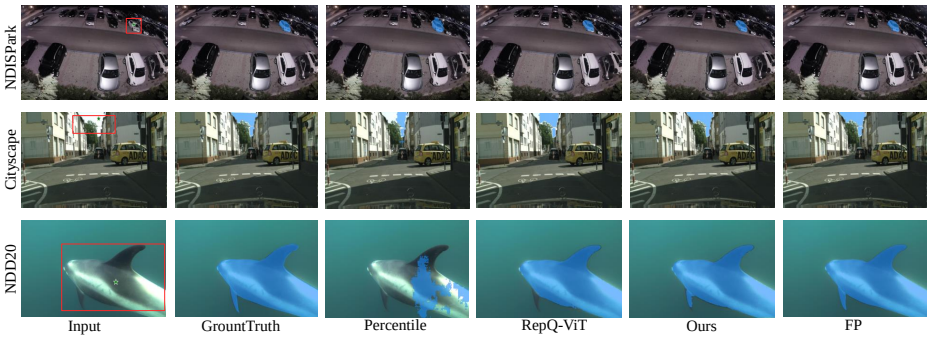

Fig. 1: Visual comparison of different PTQ methods for SAM's point prompt mode. The red block and green star represent the object to be segmented and the position of the prompt point, respectively.

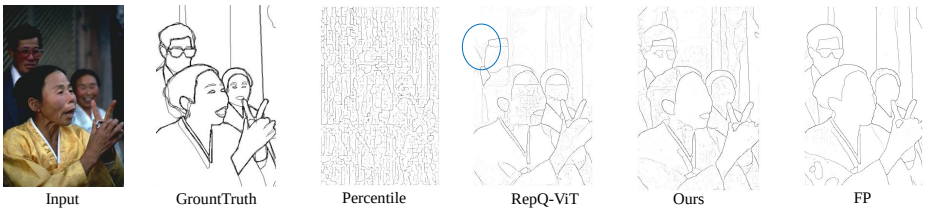

Fig. 2: Visual comparison of different PTQ methods for SAM's automatic mode. The blue circle indicates segmentation errors.

In the point prompt mode, PQ-SAM effectively preserves SAM's understanding ability of the prompt guidance, resulting in accurate segmentation masks for the specific object. Similarly, in the automatic mode, PQ-SAM excels at preserving SAM's accurate segmentation capacity, leading to high-fidelity object boundaries. This further emphasizes the effectiveness of PQ-SAM in maintaining the precision and quality of the segmentation process.

We further give more qualitative results from SAM's bounding-box prompt mode in Figure 3.

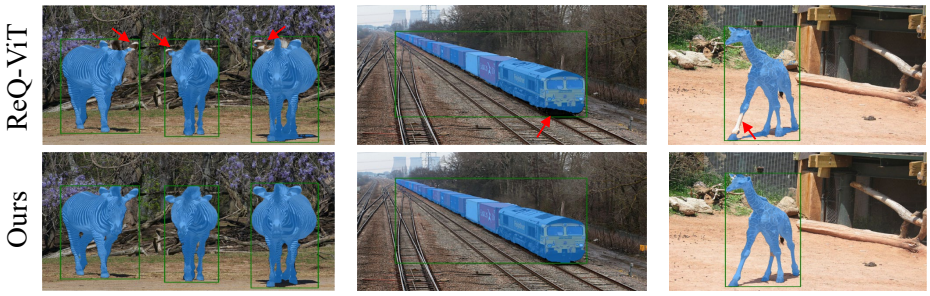

Fig. 3: Visual comparison of different PTQ methods for SAM's bounding-box mode. The green circle indicates the bounding box prompts.

3 Analysis on the activation distribution before and after GADT

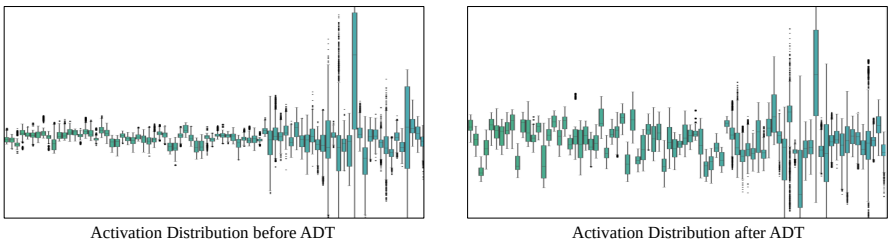

Fig. 4: Comparison of activation distributions before and after transformation by the proposed grouped activation distribution transformation (GADT).

As depicted in Figure 4, we present a visual comparison of the activation distribution for the output of the same transformer block of SAM's ViT-L version before and after applying the proposed grouped activation distribution transformation (GADT) technique based on the learned scaling and shifting sizes.

Prior to GADT, the activation values in each channel exhibited significant variations in their value ranges. However, after applying the GADT technique, the distribution of values in each channel becomes closer, which is friendly to the per-tensor distribution. More specifically, channels with small value ranges are linearly enlarged, while channels with larger value ranges are also linearly reduced.