IPTQ-ViT: Post-Training Quantization of Non-linear Functions for Integer-only Vision Transformers - Supplementary Material -

A. IPTQ-ViT Algorithm

Algorithm 1 summarizes the full pipeline of IPTQ-ViT to support implementation and clarify the method described in the main paper. To compute *Unified Metric* for a specific layer, we follow prior works [2, 3, 12] and assume that the metric can be computed independently per layer. During this process, only one activation layer is quantized at a time using each candidate approximation function. If there are N candidates, we compute N *Unified Metric* values per activation layer (e.g., Softmax, GELU, LayerNorm), enabling layer-wise selection based on sensitivity, perturbation and efficiency. The runtime of our IPTQ-ViT pipeline (including analysis, assignment and PTQ calibration) is discussed in Section E and the main paper.

B. Layer-wise Quantization Sensitivity

We apply IPTQ-ViT to ViT-B and measure the Signal-to-Quantization-Noise Ratio (SQNR) for each activation layer, where higher SQNR indicates lower quantization error. As shown in Fig. 1, IPTQ-ViT consistently achieves higher SQNR than FQ-ViT [10] and QAT-based approximations (I-ViT* and I-BERT*). Notably, I-ViT* and I-BERT* exhibit severe quantization error in GELU layers, leading to error accumulation across the model. In contrast, IPTQ-ViT maintains stable quantization quality across all layers, demonstrating the effectiveness of our approach.

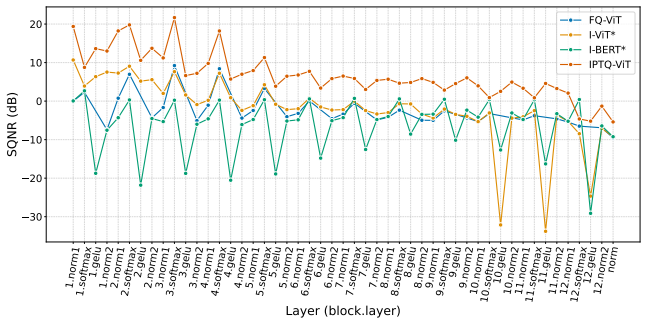


Figure 1. Visualization of layer-wise quantization sensitivity (SQNR in dB, ↑ better) of IPTQ-ViT, FQ-ViT [10], QAT-based approximation functions on ViT-B under PTQ (W4A8).

Algorithm 1: IPTQ-ViT Pipeline

Input: Calibration dataset $\mathcal{D}_{\text{calib}}$, Analysis dataset $\mathcal{D}_{\text{analysis}}$, Approximation function search space $\mathcal{S}$
Data: `metric`, which is stored evaluation metric $\Omega_f^{a_i}$ for each function-layer pair
Output: Fully Quantized Vision Transformer Q
foreach $X_a \in \mathcal{D}_{analysis}$ **do**
  Input X_a into the full-precision model to extract activations;
`/* Stage 1: Layer-wise Unified Metric Analysis */`
foreach *approximation function* $f \in \mathcal{S}$ **do**
  foreach $a_i \in$ *ViT i-th activation layers with name* `layer_name` **do**
    Replace activation layer a_i with f, resulting in quantized submodel Q_{a_i};
    `/* Calibration for` Q_{a_i} `*/`
    foreach $X_c \in \mathcal{D}_{calib}$ **do**
      Input X_c into Q_{a_i};
      Update quantization parameters of a_i;
    `/* Metric Calculation */`
    Input X_a into Q_{a_i} using f;
    Compute metric $\Omega_f^{a_i}$;
    Update `metric`$[f][$`layer_name`$] \leftarrow \Omega_f^{a_i}$;
`/* Stage 2: Approximation Function Assignment */`
foreach *activation layer* `layer_name` **do**
  Select $f' = \arg\max_f$ `metric`$[f][$`layer_name`$]$;
  Assign f' to layer `layer_name`;
`/* Stage 3: PTQ Calibration */`
foreach $X_c \in \mathcal{D}_{calib}$ **do**
  Input X_c into the quantized model Q;
  Update all quantization parameters;
return *Final quantized model* Q

C. Image Classification

C.1. Evaluation on Calibration Dataset Size

In this section, we evaluate the calibration dataset sizes of data-driven PTQ methods [9, 10, 13] and IPTQ-ViT on ImageNet-1k. In Tab. 1, FQ-ViT [10] requires 1000 images, while RepQ-ViT [9] and PTQ4ViT [13] use 32. IPTQ-ViT achieves effective quantization with as few as 4 to 8 images in both W8A8 and W4A8. IPTQ-ViT outperforms FQ-ViT [10] and RepQ-ViT [9] on all models except tiny models, using fewer calibration samples. The partial use of

floating-point operations in FQ-ViT [10] and RepQ-ViT [9] helps mitigate quantization errors in non-linear layers, offering a relative advantage on tiny models. However, IPTQ-ViT maintains competitive accuracy while supporting fully integer-only inference.

Method	Data Type	Quant Type	Data	Images
PSAQ-ViT V1 [7]	Partial-INT	DF	Synthetic	32
PSAQ-ViT V2 [8]		DF	Synthetic	32
CLAMP-ViT [11]		DF	Synthetic	32
FQ-ViT [10]		DD	Real	1K
RepQ-ViT [9]		DD	Real	32
PTQ4ViT [13]		DD	Real	32
IPTQ-ViT	Full-INT	DD	Real	4∼8

Table 1. Summary of calibration dataset requirements for each method on the ImageNet-1K. "DF" indicates data-free PTQ. "DD" indicates data-driven PTQ.

C.2. Impact of Calibration Dataset Size

We evaluate the impact of calibration dataset size on the quantized model performance by gradually reducing the number of calibration images used in IPTQ-ViT. We change the number of calibration samples from 1000 to 4 under both W8A8 and W4A8, measuring the resulting accuracy across various Vision Transformer models. As shown in Tab. 2 and Tab. 3, IPTQ-ViT maintains stable accuracy with as few as 4 to 8 calibration images. Our method is distinct in calibration dataset size from previous data-driven PTQ methods [9, 10, 13], which need larger calibration datasets.

Images	DeiT-T	DeiT-S	DeiT-B	ViT-B	ViT-L	Swin-T	Swin-S
1000	71.80	79.40	81.43	83.74	85.28	80.42	82.92
800	71.69	79.42	81.46	83.84	85.24	80.43	82.92
512	71.74	79.41	81.39	83.81	85.45	80.35	82.88
256	71.78	79.49	81.45	83.90	85.44	80.54	82.86
128	71.78	79.34	81.49	83.86	85.50	80.65	82.90
64	71.87	79.46	81.50	83.84	85.42	80.71	82.98
32	71.81	79.54	81.55	84.05	85.38	80.79	82.93
16	71.73	79.50	81.67	84.00	85.38	80.86	82.94
8	71.97	79.58	81.64	83.99	85.43	80.90	83.08
4	72.10	79.76	81.84	84.19	85.51	81.09	82.94
MAX	72.10	79.76	81.84	84.19	85.51	81.09	83.08

Table 2. Top-1 accuracy (%) of IPTQ-ViT under W8A8 with various calibration dataset sizes.

D. Object Detection

We evaluate IPTQ-ViT and QAT-based approximation functions on Cascade Mask R-CNN [1] and Mask R-CNN [4] with a Swin backbone for object detection. As discussed in the motivation section, applying QAT-based approximations directly under PTQ leads to accuracy degradation in object detection as well. As shown in Tab. 4, I-ViT* and I-BERT* exhibit significant drops in accuracy under W4A8 quantization, whereas IPTQ-ViT maintains performance.

Images	DeiT-T	DeiT-S	DeiT-B	ViT-B	ViT-L	Swin-T	Swin-S
1000	66.63	76.83	80.30	81.00	84.76	78.58	81.68
800	66.78	76.93	80.28	81.00	84.82	78.61	81.73
512	66.70	76.97	80.31	80.97	84.87	78.57	81.76
256	66.90	76.99	80.52	81.28	84.86	78.70	81.71
128	66.74	76.95	80.31	81.01	84.96	78.72	81.68
64	66.79	76.86	80.37	81.08	85.03	78.75	81.70
32	66.82	76.97	80.37	81.48	84.87	78.82	81.66
16	66.78	77.04	80.49	81.49	84.91	78.92	81.85
8	66.80	77.14	80.98	81.68	84.91	79.09	81.91
4	66.87	77.28	80.56	82.03	84.91	79.13	82.53
MAX	66.90	77.28	80.98	82.03	85.03	79.13	82.53

Table 3. Top-1 accuracy (%) of IPTQ-ViT under W4A8 with various calibration dataset sizes.

Method	W8A8		W4A8	
Swin-B	AP^{box}	AP^{mask}	AP^{box}	AP^{mask}
I-ViT*	0.7	0.6	0.3	0.2
I-BERT*	50.2	43.8	15.5	14.1
PTQ4ViT [13]	2	1.9	20	28.4
FQ-ViT [10]	51.1	44.3	N/A	N/A
IPTQ-ViT	51.7	44.9	45.0	39.2

Table 4. Object detection results on COCO dataset using Cascade Mask R-CNN [1] with *Swin-B* under W4A8 quantization.

D.1. Impact of Calibration Dataset Size

Following the same setting as in Section C.2, we evaluate the effect of calibration dataset size on IPTQ-ViT for object detection. Experiments are conducted on the COCO dataset using Cascade Mask R-CNN [1] with Swin-T and Swin-S backbones under W8A8 and W4A8 quantization settings. The range of calibration dataset sizes remains identical to that used for Section C.2. In Tab 5, IPTQ-ViT shows consistent robustness to calibration size in object detection task as shown in image classification task. For example, in the W8A8 setting with Swin-T, the difference between the minimum and maximum box AP is only 0.2, indicating that performance is less sensitive to calibration dataset size.

Images	W8A8				W4A8			
	Swin-T		Swin-S		Swin-T		Swin-S	
	AP^{box}	AP^{mask}	AP^{box}	AP^{mask}	AP^{box}	AP^{mask}	AP^{box}	AP^{mask}
1000	50.3	43.6	51.7	44.7	41.7	36.5	48.0	41.6
800	50.3	43.7	51.8	44.8	41.8	36.6	47.9	41.5
512	50.4	43.7	51.6	44.7	41.8	36.7	48.0	41.6
256	50.3	43.7	51.7	44.7	42.7	37.3	47.9	41.6
128	50.3	43.6	51.7	44.8	42.5	37.2	48.0	41.6
64	50.2	43.6	51.7	44.7	42.8	37.5	48.1	41.7
32	50.3	43.5	51.8	44.8	42.7	37.3	48.2	41.8
16	50.2	43.5	51.4	44.5	42.0	36.8	47.9	41.5
8	50.4	43.7	51.5	44.6	43.1	37.7	48.0	41.6
4	50.3	43.3	51.5	44.5	42.4	37.1	48.2	41.8
MAX	50.4	43.7	51.8	44.8	43.1	37.7	48.2	41.8

Table 5. Object detection results on COCO dataset using Cascade Mask R-CNN [1] with Swin-T and Swin-S under varying calibration dataset sizes.

Method	W/A	Swin-T		Swin-S	
		AP^{box}	AP^{mask}	AP^{box}	AP^{mask}
FQ-ViT [10]	8/8	45.3	41.2	48.2	42.6
IPTQ-ViT	8/8	45.9	41.5	48.2	43.1

Table 6. Comparison of IPTQ-ViT and FQ-ViT [10] on object detection with Mask R-CNN [4] on COCO dataset.

E. Quantization Runtime

While Section 5.4 of the main paper reports runtime only for DeiT-S, we extend the evaluation to various ViTs on a single NVIDIA 3090 GPU. The average runtimes are summarized in Tab. 7. Fig. 2 compares the runtime and top-1 accuracy of IPTQ-ViT with previous PTQ methods [7, 10, 11, 13] on ViT-B. In Fig. 2, we achieve shorter runtime than PTQ4ViT [13], PSAQ-ViT V1 [7], and CLAMP-ViT [11]. While the runtime of IPTQ-ViT is longer than FQ-ViT [10], it achieves 0.49%p higher accuracy, consistent with DeiT-S results reported in the main paper.

	DeiT-T	DeiT-S	ViT-B	Swin-T	Swin-S
Runtime (m)	2.07	2.37	3.24	2.71	4.58

Table 7. Runtime (in minutes) of IPTQ-ViT on a single NVIDIA 3090 GPU. Reported times include analysis, assignment and calibration, excluding inference.

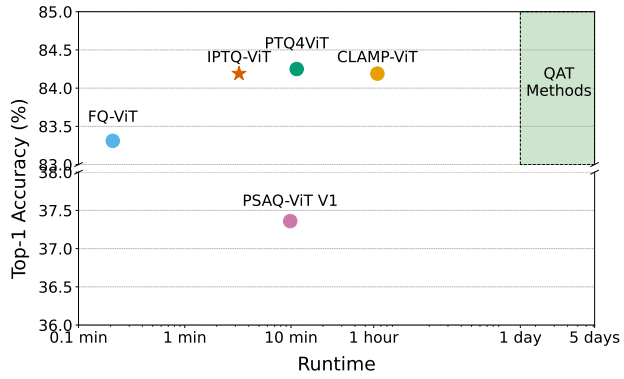


Figure 2. Runtime comparison of PTQ methods on *ViT-B* under W8A8 quantization. Reported times include quantization stages such as analysis, assignment, and calibration, excluding inference.

F. Model Size

As shown in Tab. 8, IPTQ-ViT achieves higher top-1 accuracy than PSAQ-ViT V1 [7] and V2 [8], while maintaining a similar model size. We compared the methods that support official model sizes.

Method	DeiT-T		DeiT-S		ViT-B		Swin-T		Swin-S	
	Size	Acc.	Size	Acc.	Size	Acc.	Size	Acc.	Size	Acc.
Baseline (FP)	20	72.21	88	79.85	344	84.53	116	81.35	200	83.20
PSAQ-ViT V1 [7]	2.5	71.56	11	76.92	43	37.36	14.5	75.35	25	76.64
PSAQ-ViT V2 [8]	2.5	72.17	11	79.56	43	N/A	14.5	80.20	25	82.13
IPTQ-ViT	2.5	72.10	11	79.76	43	84.19	14.5	81.09	25	83.08

Table 8. Comparison of model size (in MB) and ImageNet-1K Top-1 accuracy (%) across quantization methods [7, 8].

G. Evaluation Recipe for Latency Evaluation

All models (FP32, I-ViT [6], IPTQ-ViT) were compiled and executed with TVM on a single NVIDIA RTX 3090 GPU. For reproducibility, we provide the evaluation recipe:

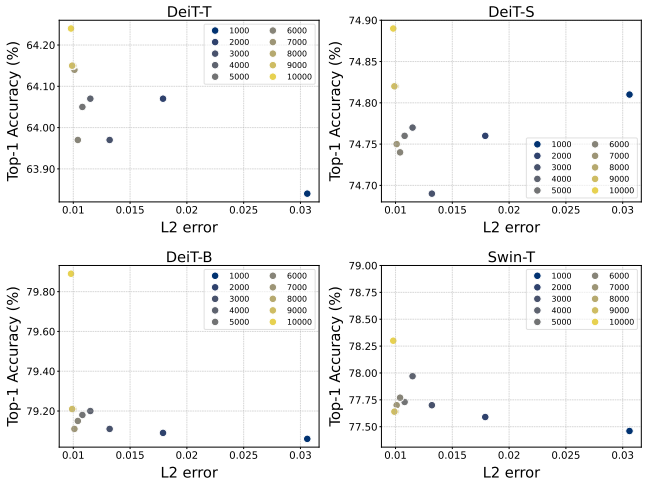
- TVM version: v0.9.0
- Batch size: 8
- Auto-tuning trials: 10,000
- Repetitions: 1,000
- Minimum repeat duration (ms): 500
- Timeout: 1,000 seconds maximum per measurement.

H. Additional Ablation Studies

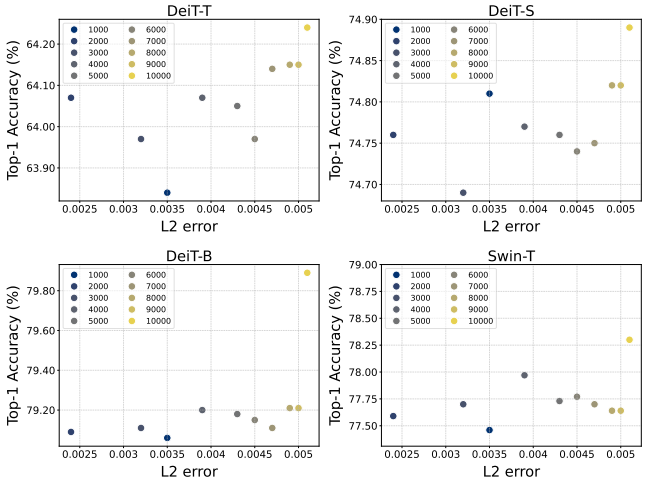
H.1. Optimization Dataset Size for Data-aware Poly-GELU

In Section 4.3 (Eq. 7) of the main paper, we compute the coefficients of *Data-aware Poly-GELU* based on the size of the optimization dataset (N). In this section, we further analyze how N affects the accuracy of our approximation. Specifically, we vary N from 1000 to 10000 and measure the L^2 errors of the erf and GELU functions, along with the ImageNet-1K top-1 accuracy under W4A8 quantization. For top-1 accuracy evaluation, it is difficult to isolate the effect of *Data-aware Poly-GELU* within the IPTQ-ViT pipeline, as it includes components such as the *Unified Metric* and *Efficient Bit-Softmax*. Therefore, following the setting from the main paper (Section 4.3), we evaluate our method by replacing i-GELU [5] with *Data-aware Poly-GELU* in the PTQ-applied I-BERT*. All other components remain unchanged, allowing us to independently assess the effect of dataset size on our proposed approximation.

As the size of the optimization dataset increases, overfitting to the GELU function can occur. However, unlike I-BERT [5], which directly fits the entire GELU, our method simplifies the optimization by indirectly fitting its core component, erf, thus improving generalization. For example, in the Swin-T, Fig.3a shows that the L^2 error of erf converges toward 0.01 with increasing dataset size, while Fig.3b shows that although the L^2 error of GELU increases, the top-1 accuracy also improves steadily. Across all evaluated models, selecting polynomial coefficients based on lower erf error consistently leads to higher top-1 accuracy than selection based on GELU error. Additionally, while significant accuracy degradation was observed under W4A8



(a) L^2 error of Eq. (6) (erf)



(b) L^2 error of Eq. (8) (GELU)

Figure 3. L^2 error of erf and GELU approximations and corresponding Top-1 accuracy (%) on ImageNet-1K, evaluated with different optimization dataset sizes (from 1000 to 10000 samples) used in Eq. (7). (a) shows the error from Eq. (6) for erf, while (b) shows the error from Eq. (8) for GELU.

when using i-GELU [5] in the motivation section of the main paper (Section. 3), replacing it with our proposed method mitigates accuracy degradation.

H.2. Effectiveness of Data-aware Poly-GELU

This section compares the proposed *Data-aware Poly-GELU* with i-GELU of I-BERT [5] in PTQ setting. All models share the same network architecture and experimental setup and apply identical approximation functions for Softmax and LayerNorm as proposed in I-BERT [5]. The only difference is the approximation used for GELU layers: either i-GELU [5] or the proposed method.

Tab. 11 reports top-1 classification accuracy for ImageNet-1K with W4A8. I-BERT*, which uses i-GELU [5], yields less than 0.1% accuracy across all models. I-BERT†, using *Data-aware Poly-GELU* instead of i-GELU [5], achieves an improvement of over 60%p accuracy under the same conditions. This substantial difference, resulting from a single function replacement, underscores the effectiveness of the proposed approximation.

We analyze quantization sensitivity using layer-wise SQNR, as shown in Tab. 9 and Tab. 10. I-BERT* exhibits a sharp decline in SQNR in deeper layers, indicating high sensitivity to quantization. I-BERT† shows enhanced SQNR across all blocks, indicating robustness. These results suggest that the proposed method provides a more precise GELU approximation than i-GELU [5] for vision data under PTQ constraints.

H.3. Micro Ablation Study on Data-aware Poly-GELU

We assess the contribution of three components of *Data-aware Poly-GELU*: (i) VR—an optimization interval derived from vision data statistics, (ii) D4—a degree-4 polynomial approximation, and (iii) erf—an erf-aligned objective. We conduct a micro ablation by replacing I-BERT*'s i-GELU [5] (in a PTQ setting) with approximations that include each component individually and in combination.

As shown in Table 15a, under W8A8 each component alone improves accuracy, and combining all three reports the largest gain, indicating that VR and erf complement D4. Under W4A8 in Table 15b, D4 is the dominant factor: replacing i-GELU [5] with D4 raises top-1 accuracy from as low as 0.08% to a level comparable to other PTQ methods. While VR and erf bring little gain on their own at W4A8, adding them on top of D4 provides an additional +0.91%p on average (up to +2.25%p on DeiT-T).

H.4. Effectiveness of Efficient Bit-Softmax

This section evaluates the effectiveness of *Efficient Bit-Softmax* in PTQ by comparing it with Shiftmax [6]. Following the same experimental setup as Section H.2, both models share the same network architecture and conditions. All non-linear functions except Softmax are approximated using the same functions proposed in I-ViT [6].

In Motivation section of the main paper, we show layer-wise SQNR of Softmax layers in ViT-B using Shiftmax [6] (I-ViT*) and our method (I-ViT†). Our method enhances SQNR across all layers. As shown in Tab. 9 and Tab. 10, I-ViT* exhibits a significant SQNR drop in deeper layers, which indicates that Shiftmax accumulates quantization errors and difficulties to remain stable in PTQ without retraining. On the other hand, I-ViT† shows improved SQNR for each Softmax layer for both DeiT-B, ViT-B and Swin-S. It demonstrates that our method provides more precise quantization performance than QAT-based method of I-ViT [6].

Method	Model	blk1	blk2	blk3	blk4	blk5	blk6	blk7	blk8	blk9	blk10	blk11	blk12
DeiT-B	I-BERT*	-32.36 (+27.29)	-59.17 (+54.44)	-47.91 (+44.07)	-41.8 (+37.8)	-37.58 (+33.06)	-37.31 (+32.36)	-34.09 (+28.82)	-31.72 (+26.09)	-37.9 (+31.7)	-45.87 (+40.45)	-46.54 (+41.06)	-31.49 (+26.85)
	I-BERT†	-5.07	-4.73	-3.84	-4.0	-4.52	-4.95	-5.27	-5.63	-6.2	-5.42	-5.48	-4.64
ViT-B	I-BERT*	-11.25 (+8.13)	-16.0 (+13.21)	-11.44 (+8.94)	-12.13 (+9.36)	-12.4 (+8.93)	-12.47 (+7.73)	-11.35 (+6.27)	-8.98 (+5.09)	-13.43 (+10.97)	-22.62 (+19.18)	-25.08 (+24.29)	-17.81 (+12.04)
	I-BERT†	-3.12	-2.79	-2.5	-2.77	-3.47	-4.74	-5.08	-3.89	-2.46	-3.44	-0.79	-5.77
DeiT-B	I-ViT*	-18.9 (+3.18)	-26.38 (+2.66)	-16.09 (+2.76)	6.42 (+1.97)	-3.75 (+2.3)	-20.98 (+2.5)	-20.5 (+2.66)	-24.75 (+2.65)	-27.36 (+2.64)	-28.92 (+2.55)	-14.07 (+1.8)	-30.72 (+2.2)
	I-ViT†	-15.72	-23.72	-13.33	8.39	-1.45	-18.48	-17.84	-22.10	-24.72	-26.37	-12.27	-28.52
ViT-B	I-ViT*	0.5 (+1.3)	-1.1 (+1.91)	-4.15 (+1.93)	-7.5 (+2.45)	**-24.13** (+2.72)	-32.66 (+2.71)	-34.65 (+2.77)	-27.42 (+2.75)	-24.61 (+2.37)	-21.04 (+2.51)	-17.22 (+2.09)	-10.72 (+1.64)
	I-ViT†	1.8	0.81	-2.22	-5.05	**-21.41**	-29.95	-31.88	-24.67	-22.24	-18.53	-15.13	-9.07

Table 9. Layer-wise quantization sensitivity (SQNR in dB) comparison for GELU and Softmax in *DeiT-B* and *ViT-B*. * uses baseline methods (i-GELU [5] for GELU, Shiftmax [6] for Softmax) and † uses our methods. Colored values indicate SQNR gain (**green**).

Method	L1.B1	L1.B2	L2.B1	L2.B2	L3.B1	L3.B2	L3.B3	L3.B4	L3.B5	L3.B6	L3.B7	L3.B8
I-BERT*	-35.38 (+27.78)	-30.97 (+21.79)	-18.32 (+13.33)	-20.87 (+15.92)	-14.43 (+13.25)	-26.26 (+23.82)	-28.04 (+25.00)	-22.09 (+20.30)	-24.12 (+21.73)	-24.27 (+21.74)	-24.74 (+21.66)	-25.03 (+21.39)
I-BERT†	-7.60	-9.18	-4.99	-4.95	-1.18	-2.44	-3.04	-1.79	-2.38	-2.53	-3.08	-3.64
I-ViT*	-26.92 (+1.68)	-26.80 (+1.70)	-12.22 (+1.63)	-2.91 (+1.24)	-16.55 (+1.61)	-6.50 (+1.45)	-24.21 (+1.69)	-16.42 (+1.32)	-29.72 (+1.57)	-16.38 (+1.19)	-26.64 (+1.33)	-18.13 (+1.19)
I-ViT†	-25.24	-25.10	-10.59	-1.67	-14.94	-5.05	-22.52	-15.10	-28.15	-15.19	-25.31	-16.94
Method	**L3.B9**	**L3.B10**	**L3.B11**	**L3.B12**	**L3.B13**	**L3.B14**	**L3.B15**	**L3.B16**	**L3.B17**	**L4.B1**	**L4.B2**	
I-BERT*	-26.27 (+21.96)	-27.10 (+23.32)	-33.64 (+28.86)	-32.14 (+28.43)	-34.39 (+30.45)	-28.90 (+25.03)	-35.41 (+29.95)	-39.29 (+33.39)	-75.15 (+62.90)	-68.67 (+50.68)	-22.26 (+21.89)	
I-BERT†	-4.32	-3.78	-4.77	-3.71	-3.94	-3.87	-5.46	-5.90	-12.25	-17.99	-0.38	
I-ViT*	-25.71 (+1.24)	-7.71 (+0.71)	-13.65 (+0.97)	-4.82 (+0.37)	-14.83 (+0.92)	-6.06 (+0.34)	-10.42 (+0.70)	-4.24 (-0.12)	-7.57 (-0.20)	-5.66 (-0.13)	-4.21 (+0.41)	
I-ViT†	-24.47	-7.00	-12.68	-4.45	-13.91	-5.72	-9.72	-4.36	-7.77	-5.79	-3.80	

Table 10. Layer-wise quantization sensitivity (SQNR in dB) comparison for GELU and Softmax in *Swin-S*. ∗ uses baseline methods (i-GELU [5] for GELU, Shiftmax [6] for Softmax) and † uses our methods. Colored values indicate SQNR gain (**green**) or degradation (**red**).

Method	DeiT-T	DeiT-S	DeiT-B	ViT-B	ViT-L	Swin-T	Swin-S
I-BERT*	0.08	0.10	0.10	0.09	0.10	0.10	0.10
I-BERT†	64.24	74.89	79.89	63.70	82.47	78.29	81.62

Table 11. Top-1 accuracy (%) of polynomial approximation functions on ImageNet-1k. * means quantized with I-BERT [5] methods. † means the same quantization as *, except for GELU, which is quantized with ours. All models are quantized with W4A8.

I. Layer-wise Assignment Counts of Approximation Functions

Table 12 summarizes, for each non-linear layer, how often each method is selected after the IPTQ-ViT pipeline.

J. Efficient Bit-Softmax: Trade-off Analysis and Calibration Overfitting

In this section, we analyze why the degree-1 (D1) of Eq. (13) outperforms the degree-2 (D2) in Sec. 4.4 of main paper. For each model and bit-width, we measure layer-wise Softmax SQNR on the calibration and validation sets and compute the gap $\Delta SQNR = SQNR_{val} - SQNR_{calib}$. Results are summarized in Table 14 and Fig. 5 (with $\Delta SQNR$ shown to the right of the "val" value and above the "bar" plot). D2 shows consistently larger negative $\Delta SQNR$. For example, on DeiT-T with W8A8, D1 shows an average SQNR increase of +1.52 dB, while D2 shows a drop of 3.02 dB (Table 14). Similarly, in W4A8, they show a drop of 1.44 dB, while D1 shows an increase of +0.08 dB. The same pattern is observed in Swin-T and Swin-S, aligning with the results in Tab. 4 of main paper.

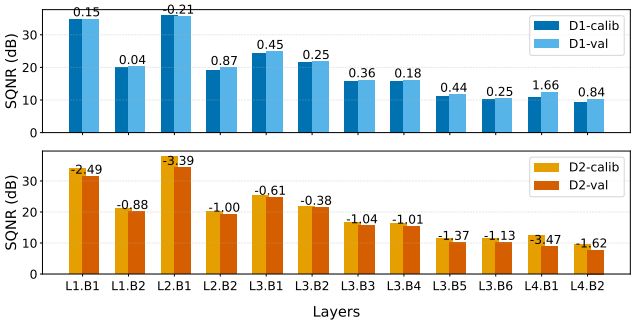


Figure 5. Calibration vs. validation SQNR on Swin-T W4A8. Top: Ours-D1, showing stable generalization. Bottom: Ours-D2, showing overfitting. Numbers above bars: $SQNR_{\text{val}} - SQNR_{\text{calib}}$.

Model	Softmax				GELU			LayerNorm		
	Ours	I-ViT	FQ-ViT	I-BERT	Ours	I-ViT	I-BERT	I-ViT	FQ-ViT	I-BERT
DeiT-T	6	0	4	2	9	2	1	4	6	15
DeiT-S	7	3	1	1	10	1	1	2	6	17
DeiT-B	4	2	6	0	7	4	1	6	3	16
ViT-B	3	0	8	1	9	2	1	4	12	9
Swin-T	5	2	2	3	8	1	3	9	6	10
Swin-S	17	4	2	1	18	6	0	18	19	12

Table 12. Layer-wise assignment counts of approximation functions after the IPTQ-ViT pipeline.

Method	L^2 error	L^∞ error
Eq.10 (I-ViT)	0.0454	0.0990
Eq.13 (Ours)	0.0347	0.0430

Table 13. Comparison of Eq. (10) and Eq. (13) in the main paper over the range $(-\ln 2, 0)$.

W/A	Degree	Type	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12
8/8	Ours-D1	calib	49.52	35.06	33.49	34.50	28.55	35.86	23.95	25.76	25.43	26.94	21.90	13.95
8/8	Ours-D1	val	52.24 (+2.72)	38.26 (+3.20)	40.17 (+6.68)	36.24 (+1.74)	29.44 (+0.89)	33.89 (-1.97)	26.03 (+2.08)	23.53 (-2.23)	27.02 (+1.59)	27.77 (+0.83)	25.16 (+3.26)	13.37 (-0.58)
8/8	Ours-D2	calib	58.24	28.60	34.68	35.91	28.07	29.47	23.25	25.88	26.42	26.66	25.41	13.62
8/8	Ours-D2	val	49.21 (-9.03)	25.98 (-2.62)	30.23 (-4.45)	28.98 (-6.93)	25.73 (-2.34)	27.96 (-1.51)	22.78 (-0.47)	23.40 (-2.48)	24.88 (-1.54)	24.23 (-2.43)	24.10 (-1.31)	12.48 (-1.14)
4/8	Ours-D1	calib	35.87	24.84	22.06	20.31	19.80	18.55	12.58	12.83	13.80	13.39	12.96	3.28
4/8	Ours-D1	val	35.36 (-0.51)	24.77 (-0.07)	22.36 (+0.30)	20.47 (+0.16)	19.21 (-0.59)	18.57 (+0.02)	12.78 (+0.20)	13.24 (+0.41)	13.68 (-0.12)	13.30 (-0.09)	13.46 (+0.50)	3.99 (+0.71)
4/8	Ours-D2	calib	33.67	20.18	19.31	18.23	13.71	16.00	13.08	14.72	14.46	13.28	13.53	6.40
4/8	Ours-D2	val	33.91 (+0.24)	20.07 (-0.11)	19.12 (-0.19)	17.48 (-0.75)	12.75 (-0.96)	14.36 (-1.64)	11.30 (-1.78)	12.12 (-2.60)	12.87 (-1.59)	11.45 (-1.83)	11.47 (-2.06)	2.43 (-3.97)

(a) DeiT-T

W/A	Degree	Type	L1.B1	L1.B2	L2.B1	L2.B2	L3.B1	L3.B2	L3.B3	L3.B4	L3.B5	L3.B6	L4.B1	L4.B2
8/8	Ours-D1	calib	41.59	28.27	44.13	26.32	31.81	28.27	22.62	22.51	17.22	15.24	15.80	13.69
8/8	Ours-D1	val	40.98 (-0.61)	30.48 (+2.21)	46.60 (+2.47)	28.55 (+2.23)	32.80 (+0.99)	29.49 (+1.22)	24.26 (+1.64)	24.10 (+1.59)	19.99 (+2.77)	18.30 (+3.06)	18.38 (+2.58)	17.05 (+3.36)
8/8	Ours-D2	calib	44.22	28.82	45.32	26.40	30.99	28.48	23.89	23.57	18.13	16.59	16.30	14.92
8/8	Ours-D2	val	39.50 (-4.72)	28.11 (-0.71)	43.26 (-2.06)	26.26 (-0.14)	29.69 (-1.30)	27.47 (-1.01)	22.29 (-1.60)	22.51 (-1.06)	17.08 (-1.05)	15.77 (-0.82)	15.83 (-0.47)	14.69 (-0.23)
4/8	Ours-D1	calib	34.75	20.21	35.92	19.10	24.45	21.55	15.85	15.95	11.31	10.24	10.83	9.49
4/8	Ours-D1	val	34.91 (+0.16)	20.25 (+0.04)	35.71 (-0.21)	19.96 (+0.86)	24.91 (+0.46)	21.80 (+0.25)	16.21 (+0.36)	16.13 (+0.18)	11.75 (+0.44)	10.49 (+0.25)	12.49 (+1.66)	10.33 (+0.84)
4/8	Ours-D2	calib	34.01	21.03	37.76	20.30	25.35	21.89	16.59	16.43	11.57	11.35	12.36	9.37
4/8	Ours-D2	val	31.53 (-2.48)	20.15 (-0.88)	34.37 (-3.39)	19.30 (-1.00)	24.75 (-0.60)	21.51 (-0.38)	15.55 (-1.04)	15.43 (-1.00)	10.20 (-1.37)	10.22 (-1.13)	8.90 (-3.46)	7.75 (-1.62)

(b) Swin-T

Bit	Degree	Type	L1.B1	L1.B2	L2.B1	L2.B2	L3.B1	L3.B2	L3.B3	L3.B4	L3.B5	L3.B6	L3.B7	L3.B8
8	Ours-D1	calib	39.15	32.04	44.14	30.95	35.63	35.61	28.22	26.80	22.50	21.78	17.57	14.50
8	Ours-D1	val	39.24 (+0.09)	33.86 (+1.82)	46.41 (+2.27)	34.27 (+3.32)	38.36 (+2.73)	38.36 (+2.75)	31.06 (+2.84)	30.40 (+3.60)	25.53 (+3.03)	24.73 (+2.95)	19.85 (+2.28)	17.51 (+3.01)
8	Ours-D2	calib	40.22	33.93	43.83	30.21	33.52	38.09	29.65	28.11	23.72	23.66	18.98	16.51
8	Ours-D2	val	43.15 (+2.93)	30.73 (-3.20)	42.88 (-0.95)	31.08 (+0.87)	35.34 (+1.82)	33.56 (-4.53)	27.90 (-1.75)	27.53 (-0.58)	22.64 (-1.08)	21.93 (-1.73)	17.87 (-1.11)	15.23 (-1.28)
4	Ours-D1	calib	32.67	23.69	37.01	25.19	31.84	30.47	23.71	22.05	16.97	16.67	12.81	10.73
4	Ours-D1	val	33.16 (+0.49)	24.61 (+0.92)	39.08 (+2.07)	27.05 (+1.86)	32.78 (+0.94)	31.84 (+1.37)	24.29 (+0.58)	23.15 (+1.10)	18.09 (+1.12)	18.49 (+1.82)	13.67 (+0.86)	12.48 (+1.75)
4	Ours-D2	calib	35.98	24.56	37.22	25.53	29.34	30.02	22.36	21.29	16.87	17.29	12.86	11.67
4	Ours-D2	val	38.08 (+2.10)	23.90 (-0.66)	36.46 (-0.76)	24.80 (-0.73)	32.09 (+2.75)	30.16 (+0.14)	22.67 (+0.31)	21.68 (+0.39)	16.71 (-0.16)	15.77 (-1.52)	11.90 (-0.96)	10.15 (-1.52)

Bit	Degree	Type	L3.B9	L3.B10	L3.B11	L3.B12	L3.B13	L3.B14	L3.B15	L3.B16	L3.B17	L3.B18	L4.B1	L4.B2
8	Ours-D1	calib	11.87	9.72	7.92	6.23	6.64	6.17	6.51	5.84	5.49	5.78	16.94	14.89
8	Ours-D1	val	13.23 (+1.36)	12.73 (+3.01)	9.78 (+1.86)	9.05 (+2.82)	9.47 (+2.83)	9.15 (+2.98)	9.23 (+2.72)	8.49 (+2.65)	8.36 (+2.87)	8.53 (+2.75)	15.50 (-1.44)	13.12 (-1.77)
8	Ours-D2	calib	12.62	10.88	8.05	7.02	7.76	6.89	7.12	7.13	6.70	7.11	16.12	13.86
8	Ours-D2	val	11.67 (-0.95)	10.09 (-0.79)	7.06 (-0.99)	6.43 (-0.59)	6.45 (-1.31)	6.13 (-0.76)	6.20 (-0.92)	5.79 (-1.34)	6.00 (-0.70)	5.97 (-1.14)	15.09 (-1.03)	13.31 (-0.55)
4	Ours-D1	calib	7.44	6.00	3.50	2.40	2.39	1.67	2.15	1.28	1.56	1.65	12.16	8.54
4	Ours-D1	val	9.27 (+1.83)	7.94 (+1.94)	4.96 (+1.46)	4.81 (+2.41)	4.13 (+1.74)	4.91 (+3.24)	3.99 (+1.84)	4.35 (+3.07)	3.30 (+1.74)	4.58 (+2.93)	12.50 (+0.34)	11.47 (+2.93)
4	Ours-D2	calib	8.66	7.64	5.74	4.25	4.60	4.51	4.56	4.23	3.94	3.87	8.18	8.34
4	Ours-D2	val	7.02 (-1.64)	4.66 (-2.98)	1.29 (-4.45)	0.77 (-3.48)	0.10 (-4.50)	-0.11 (-4.62)	-0.53 (-5.09)	-0.27 (-4.50)	-0.66 (-4.60)	0.38 (-3.49)	11.04 (+2.86)	7.39 (-0.95)

(c) Swin-S

Table 14. Layer-wise SQNR (dB) of *Efficient Bit-Softmax* on **DeiT-T**, **Swin-T**, and **Swin-S** for degree-1 (D1) and degree-2 (D2) polynomials in Eq. (13) of the main paper. D2 shows a larger drop from calibration to validation, indicating calibration overfitting. $\Delta\text{SQNR} = \text{SQNR}_{\text{val}} - \text{SQNR}_{\text{calib}}$; **green**/**red** mark positive/negative ΔSQNR.

VR	D4	erf	DeiT-T	DeiT-S	DeiT-B	Swin-T	Swin-S
*I-BERT**			68.37	77.31	80.88	35.00	82.49
✓			68.30	77.42	80.81	35.23	82.51
	✓		68.52	77.72	80.95	80.22	82.67
		✓	67.71	77.51	80.78	36.00	82.55
✓	✓		68.52	77.65	81.07	80.23	82.69
	✓	✓	68.68	77.77	81.02	80.27	82.71
✓		✓	68.24	77.55	80.72	42.06	82.55
✓	✓	✓	68.73	77.76	81.07	80.24	82.75

(a) W8A8

VR	D4	erf	DeiT-T	DeiT-S	DeiT-B	Swin-T	Swin-S
*I-BERT**			0.08	0.10	0.10	0.10	0.10
✓			0.09	0.10	0.10	0.10	0.11
	✓		61.99	73.99	78.98	78.11	81.32
		✓	0.09	0.10	0.10	0.11	0.10
✓	✓		62.00	73.99	79.23	78.31	81.32
	✓	✓	62.67	74.23	79.67	78.06	81.46
✓		✓	0.10	0.10	0.10	0.12	0.10
✓	✓	✓	64.24	74.89	79.89	78.29	81.62

(b) W4A8

Table 15. Micro ablation study of *Data-aware Poly-GELU* on ImageNet-1K under **W8A8** (left) and **W4A8** (right). We analyze the impact of three components: Vision data input Range for optimizing approximation function (VR), Degree-4 polynomial (D4), and erf-based optimization (erf). A ✓ denotes the use of a component, and I-BERT* is the baseline.

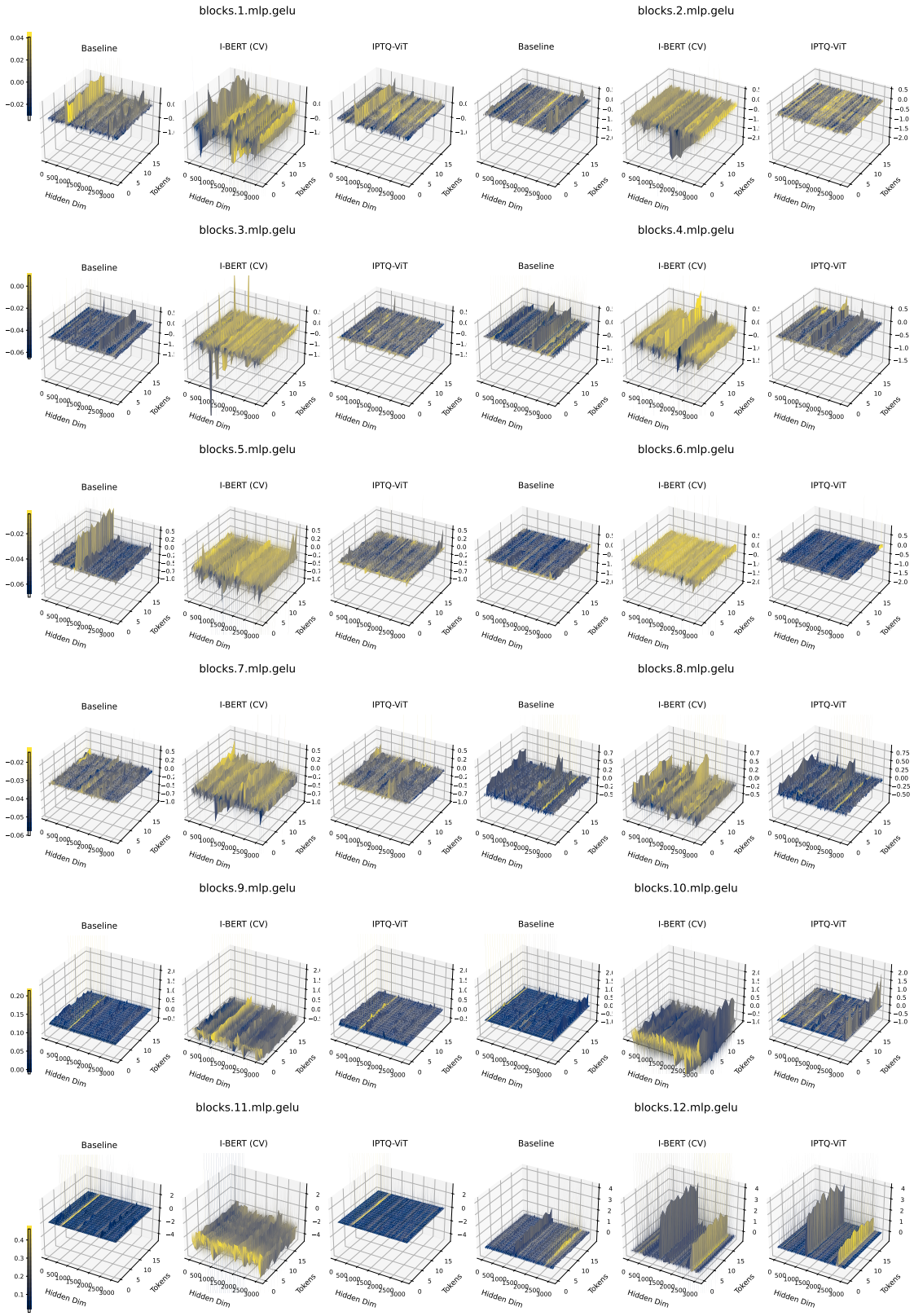


Figure 4. Extended experiments for the main paper (Section 3, Fig. 1): Comparison of activation distributions across all GELU layers in *ViT-B* between the full-precision model, i-GELU [5], and our proposed approximation (*Data-aware Poly-GELU*). I-BERT (CV), quantized under PTQ, exhibits the same degradation described in Section 3. In contrast, our method better preserves the original activation distribution.

References

[1] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018. 2

[2] Zhen Dong, Zhewei Yao, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Hawq: Hessian aware quantization of neural networks with mixed-precision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 293–302, 2019. 1

[3] Zhen Dong, Zhewei Yao, Daiyaan Arfeen, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Hawq-v2: Hessian aware trace-weighted quantization of neural networks. *Advances in neural information processing systems*, 33:18518–18529, 2020. 1

[4] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2, 3

[5] Sehoon Kim, Amir Gholami, Zhewei Yao, Michael W Mahoney, and Kurt Keutzer. I-bert: Integer-only bert quantization. In *International conference on machine learning*, pages 5506–5518. PMLR, 2021. 3, 4, 5, 7

[6] Zhikai Li and Qingyi Gu. I-vit: Integer-only quantization for efficient vision transformer inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17065–17075, 2023. 3, 4, 5

[7] Zhikai Li, Liping Ma, Mengjuan Chen, Junrui Xiao, and Qingyi Gu. Patch similarity aware data-free quantization for vision transformers. In *European conference on computer vision*, pages 154–170. Springer, 2022. 2, 3

[8] Zhikai Li, Mengjuan Chen, Junrui Xiao, and Qingyi Gu. Psaq-vit v2: Toward accurate and general data-free quantization for vision transformers. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 2, 3

[9] Zhikai Li, Junrui Xiao, Lianwei Yang, and Qingyi Gu. Repq-vit: Scale reparameterization for post-training quantization of vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17227–17236, 2023. 1, 2

[10] Yang Lin, Tianyu Zhang, Peiqin Sun, Zheng Li, and Shuchang Zhou. Fq-vit: Post-training quantization for fully quantized vision transformer. *arXiv preprint arXiv:2111.13824*, 2021. 1, 2, 3

[11] Akshat Ramachandran, Souvik Kundu, and Tushar Krishna. Clamp-vit: Contrastive data-free learning for adaptive post-training quantization of vits. In *European Conference on Computer Vision*, pages 307–325. Springer, 2024. 2, 3

[12] Zhewei Yao, Zhen Dong, Zhangcheng Zheng, Amir Gholami, Jiali Yu, Eric Tan, Leyuan Wang, Qijing Huang, Yida Wang, Michael Mahoney, et al. Hawq-v3: Dyadic neural network quantization. In *International Conference on Machine Learning*, pages 11875–11886. PMLR, 2021. 1

[13] Zhihang Yuan, Chenhao Xue, Yiqi Chen, Qiang Wu, and Guangyu Sun. Ptq4vit: Post-training quantization for vision transformers with twin uniform quantization. In *European conference on computer vision*, pages 191–207. Springer, 2022. 1, 2, 3