VLM-PTQ: Efficient Post-Training Quantization for Large Vision-Language Models

Juncan Deng[1] Kejie Huang[1*]
[1]Zhejiang University
{dengjuncan,huangkejie}@zju.edu.cn

1. Appendix

1.1. Proof of Precomputation

A complete proof is provided in the GPTAQ paper. Here, we briefly recapitulate it using the computation of $\mathbf{P}$ as an example. For any further questions, we refer the reader to the proof in Appendix A.3 of the GPTAQ paper. We derive $\mathbf{P}$ row-wise. Starting from the expression

$$\mathbf{P}_{i,:} = \Delta\mathbf{X}_{i,:}\mathbf{X}^\top \mathbf{L}_{i+1:,i+1:}\mathbf{L}^\top_{i+1:,i+1:}, \tag{1}$$

note that $\mathbf{L}^\top_{i+1:,i+1:}$ has support only in columns $j > i$. Thus, $\mathbf{P1}_{i,j} = 0$ for $i \geq j$, and for $i < j$,

$$\mathbf{P}_{i,j} = \sum_{a=i+1}^{j} \mathbf{O}_{i,a}\mathbf{L}^\top_{a,j}, \tag{2}$$

where $\mathbf{O}_{i,a} := (\Delta\mathbf{X}_{i,:}\mathbf{X}^\top\mathbf{L}_{i+1:,i+1:})_a$.

Since $\mathbf{L}_{i+1:,i+1:}$ is zero in columns $a \leq i$, we have $\mathbf{O}_{i,a} = 0$ for $a \leq i$, and for $a > i$,

$$\mathbf{O}_{i,a} = \sum_{b=a}^{n} (\Delta\mathbf{X}\mathbf{X}^\top)_{i,b}\,\mathbf{L}_{b,a}. \tag{3}$$

This means $\mathbf{O} = (\Delta\mathbf{X}\mathbf{X}^\top\mathbf{L}) \odot \text{mask}_\mathbf{U}$, where $\odot$ denotes element-wise multiplication and $\text{mask}_\mathbf{U}$ masks out the lower triangle (including diagonal).

Finally, since $\mathbf{O}$ is strictly upper-triangular, multiplying by $\mathbf{L}^\top$ yields:

$$(\mathbf{O}\mathbf{L}^\top)_{i,j} = \sum_{a=1}^{n} \mathbf{O}_{i,a}\mathbf{L}^\top_{a,j} = \sum_{a=i+1}^{j} \mathbf{O}_{i,a}\mathbf{L}^\top_{a,j} = \mathbf{P}_{i,j}, \tag{4}$$

because $\mathbf{O}_{i,a} = 0$ for $a \leq i$ and $\mathbf{L}^\top_{a,j} = 0$ for $a > j$. Hence,

$$\mathbf{P} = \big((\Delta\mathbf{X}\mathbf{X}^\top\mathbf{L}) \odot \text{mask}_\mathbf{U}\big)\mathbf{L}^\top. \tag{5}$$

*Corresponding Authors

1.2. Comparison with Additional Competitive Methods

To further validate the effectiveness of our approach against a broader range of state-of-the-art quantization methods, we conduct comprehensive comparisons with additional competitive baselines in the 3-bit weight-only quantization setting. Specifically, we compare against AWQ [3], MBQ [2], and VLMQ [4]. Table 1 presents the detailed results across eight representative VLM models ranging from 0.5B to 72B parameters. As shown in Table 1, our method consistently outperforms all competing approaches across diverse VLM architectures and model scales. On the challenging ultra-small LLaVA-OneVision-0.5B model, where 3-bit quantization causes severe performance degradation, our method achieves 20.0% average accuracy, substantially outperforming VLMQ at 14.1%, MBQ at 12.9%, and AWQ at 11.3%, demonstrating improvements of 5.1, 7.1, and 8.7 percentage points, respectively. On the Qwen2.5-VL-7B-Instruct model, our approach achieves 71.3% average accuracy, surpassing VLMQ by 6.1 percentage points, MBQ by 7.0 percentage points, and AWQ by 9.8 percentage point. The performance gap becomes more pronounced on larger models, with our approach achieving 76.9% average accuracy on Qwen2.5-VL-72B-Instruct compared to 72.1% for VLMQ, 70.7% for MBQ, and 67.1% for AWQ. On the LLaVA-OneVision-7B model, our method achieves 69.5% average accuracy, demonstrating consistent advantages with 3.8 percentage points over VLMQ at 65.7%, 5.1 percentage points over MBQ at 64.4%, and 8.9 percentage points over AWQ at 60.6%. These consistent improvements across the entire model spectrum from 0.5B to 72B parameters, spanning both Qwen, InternVL, and LLaVA model families, demonstrate that our method establishes a new state-of-the-art for post-training quantization of vision-language models.

1.3. Extreme Low-Bit Activation Quantization

To demonstrate the robustness of our method under extremely aggressive quantization settings, we conduct experiments with 4-bit activation quantization combined with

Table 1. Comparison with additional competitive methods under 3-bit weight-only quantization. We evaluate on eight benchmarks: ChartQA, DocVQA (Validation set), MME-RealWorld (English), MME-RealWorld (Chinese), OCRBench, ScienceQA, SeedBench 2 Plus, and TextVQA (Validation set).

Model	Method	Chart	Doc^{val}	MME_{en}	MME_{cn}	OCR	SciQA	Seed^{2+}	Text^{val}	Avg (↑)
LLaVA-OneVision -0.5B	FP16	61.5	69.0	38.9	32.1	57.6	63.4	53.1	65.8	63.5
	AWQ	6.5	12.5	14.8	16.2	24.3	6.5	1.8	8.1	11.3
	MBQ	7.2	14.8	18.5	14.5	27.2	7.1	2.5	11.5	12.9
	VLMQ	8.1	16.4	20.4	15.1	29.2	7.9	3.0	12.8	14.1
	Ours	**16.2**	**22.8**	**25.6**	**19.4**	**31.5**	**14.3**	**8.2**	**21.7**	**20.0**
InternVL3-1B -Instruct	FP16	75.1	80.2	39.0	31.9	78.8	91.1	58.3	73.8	66.0
	AWQ	46.2	59.4	26.3	18.1	62.0	57.6	46.9	58.8	46.9
	MBQ	50.5	58.5	26.8	17.2	61.8	61.8	42.5	61.5	47.6
	VLMQ	51.3	59.0	27.1	18.0	62.4	62.4	43.0	62.2	48.2
	Ours	**66.6**	**72.7**	**28.0**	**19.0**	**71.7**	**82.1**	**52.0**	**69.9**	**57.8**
Qwen2.5-VL -3B-Instruct	FP16	83.3	82.9	50.1	40.4	78.2	83.4	68.1	78.9	70.6
	AWQ	70.8	75.6	31.2	30.3	64.7	63.0	59.6	63.0	57.3
	MBQ	70.5	79.8	41.5	33.2	66.2	73.5	59.8	69.5	61.8
	VLMQ	71.2	80.7	42.3	34.0	66.9	74.3	60.3	70.3	62.5
	Ours	**73.3**	**86.9**	**45.9**	**36.3**	**71.2**	**75.2**	**64.3**	**74.1**	**65.9**
Qwen2.5-VL -7B-Instruct	FP16	83.8	94.9	58.7	52.8	84.7	88.8	70.9	83.0	77.2
	AWQ	62.8	82.9	38.4	31.2	71.2	76.9	60.5	67.5	61.5
	MBQ	64.5	87.2	39.8	28.5	76.8	76.8	62.8	77.8	64.3
	VLMQ	65.2	88.0	40.8	29.1	77.5	77.7	63.8	78.7	65.2
	Ours	**69.8**	**92.3**	**51.3**	**43.8**	**80.2**	**84.1**	**67.9**	**81.4**	**71.3**
LLaVA-OneVision -7B	FP16	80.1	87.1	57.4	53.9	62.1	90.0	64.8	76.0	71.4
	AWQ	70.5	74.2	43.1	38.5	52.8	81.5	56.2	68.3	60.6
	MBQ	74.5	77.2	45.8	42.2	57.2	85.2	61.2	71.8	64.4
	VLMQ	75.6	78.3	48.9	45.4	58.0	85.3	60.9	73.1	65.7
	Ours	**77.8**	**84.5**	**54.2**	**51.8**	**60.5**	**88.2**	**63.5**	**75.2**	**69.5**
InternVL3-14B -Instruct	FP16	87.8	92.9	58.5	53.9	83.8	98.0	70.2	80.9	78.2
	AWQ	65.8	62.2	41.0	38.8	49.7	83.9	63.0	54.2	57.3
	QuaRot	72.6	82.3	41.7	37.3	74.3	85.4	63.5	72.7	66.3
	VLMQ	80.4	83.6	47.3	44.0	74.5	93.4	67.0	75.1	70.7
	Ours	**84.7**	**90.3**	**54.7**	**52.3**	**82.4**	**95.0**	**68.8**	**79.5**	**76.0**
InternVL3-38B -Instruct	FP16	88.7	93.3	61.1	60.0	85.1	98.6	71.5	83.6	80.2
	AWQ	64.8	60.0	39.4	33.3	56.8	93.2	67.7	57.6	59.1
	QuaRot	79.7	88.2	55.1	54.9	80.2	94.4	68.1	78.9	74.9
	VLMQ	84.6	84.8	46.9	41.1	80.0	95.6	68.2	81.0	72.7
	Ours	**86.7**	**90.8**	**57.4**	**57.1**	**83.2**	**97.2**	**70.4**	**82.6**	**78.2**
Qwen2.5-VL -72B-Instruct	FP16	88.4	95.6	57.5	51.5	83.8	94.1	72.7	82.0	78.2
	AWQ	72.1	81.3	43.8	37.1	75.6	86.8	63.8	75.9	67.1
	MBQ	76.5	82.8	48.2	44.5	78.5	90.2	66.5	78.5	70.7
	VLMQ	77.8	84.2	49.5	46.2	79.8	91.5	67.8	79.8	72.1
	Ours	**85.2**	**89.4**	**56.2**	**55.8**	**82.1**	**95.7**	**69.2**	**81.4**	**76.9**

4-bit weights and 8-bit KV cache (W4A4KV8). Following QuaRot [1], we first apply rotation transformations to the model weights to reduce outliers in activation distributions before quantization. Table 2 presents the results on InternVL3 models with varying scales.

As demonstrated in Table 2, even under the extremely challenging W4A4KV8 quantization setting where both weights and activations are compressed to 4 bits, our method maintains substantial performance advantages over baseline approaches. On InternVL3-1B-Instruct, our approach achieves 33.8% average accuracy, outperforming GPTAQ at 26.9% and GPTQ at 26.8% by 6.9 and 7.0 per-

Table 2. Performance under extreme low-bit quantization (W4A4KV8) with QuaRot rotation. We evaluate on eight benchmarks: ChartQA, DocVQA (Validation set), MME-RealWorld (English), MME-RealWorld (Chinese), OCRBench, ScienceQA, SeedBench 2 Plus, and TextVQA (Validation set).

Model	Method	Chart	Doc^{val}	MME_{en}	MME_{cn}	OCR	SciQA	Seed^{2+}	Text^{val}	Avg
InternVL3-1B -Instruct	FP16	75.1	80.2	39.0	31.9	78.8	91.1	58.3	73.8	66.0
	GPTQ	23.0	23.9	21.8	18.2	35.3	42.9	26.4	23.2	26.8
	GPTAQ	17.4	31.0	21.1	17.5	36.0	38.6	27.3	26.5	26.9
	Ours	**31.7**	**41.9**	**22.2**	**18.3**	**50.3**	**43.5**	**28.9**	**33.7**	**33.8**
InternVL3-14B -Instruct	FP16	87.8	92.9	58.5	53.9	83.8	98.0	70.2	80.9	78.2
	GPTQ	71.3	77.3	43.2	31.6	70.8	81.7	62.5	69.9	63.5
	GPTAQ	72.6	79.1	42.8	33.8	73.0	82.3	62.2	71.0	64.6
	Ours	**73.0**	**81.9**	**47.8**	**39.6**	**75.6**	**85.9**	**64.2**	**73.4**	**67.7**
InternVL3-38B -Instruct	FP16	88.7	93.3	61.1	60.0	85.1	98.6	71.5	83.6	80.2
	GPTQ	73.8	79.5	45.8	34.2	73.5	84.2	64.8	72.5	66.0
	GPTAQ	75.2	81.3	46.5	36.8	75.8	85.6	66.2	74.2	67.7
	Ours	**76.8**	**84.5**	**50.2**	**42.8**	**78.2**	**88.5**	**68.5**	**76.8**	**70.8**

centage points respectively. The improvements scale with model size: InternVL3-14B-Instruct reaches 67.7% with our method compared to 64.6% for GPTAQ and 63.5% for GPTQ, representing improvements of 3.1 and 4.2 percentage points, respectively. On the largest model, InternVL3-38B-Instruct, our method achieves 70.8% average accuracy, maintaining a 3.1 percentage point advantage over GPTAQ at 67.7% and a 4.8 percentage point advantage over GPTQ at 66.0%. These results demonstrate that our closed-form correction term and modality-aware quantization provide robust benefits even under extreme quantization constraints where activation quantization introduces significant additional errors. The consistent improvements across all model scales validate that our approach effectively addresses the compound challenges of simultaneous ultra-low-bit weight and activation quantization in vision-language models.

1.4. Additional Analysis on Algorithm Efficiency

To clarify the computational overhead of our method, Table 3 compares the matrix storage requirements during calibration across GPTQ, GPTAQ, and our VLM-PTQ approach. All methods share fundamental matrices for weight representation and Cholesky decomposition. The key difference lies in precomputation: GPTQ requires no additional matrices, GPTAQ introduces the correction matrix P, while our method adds the closed-form correction term C and modality-aware scaling vector M. The modest memory increase enables superior quantization quality while maintaining comparable runtime to GPTAQ.

1.5. Ablation Study on Grid Search and Calibration Samples

Our modality-aware quantization strategy requires selecting an optimal awareness coefficient μ via a lightweight grid search procedure applied to calibration data. We conduct systematic ablation experiments on the Qwen2.5-VL-7B-Instruct model under W3A16 quantization to investigate the impact of grid granularity and calibration sample quantity. The experimental results are summarized in Table 4. Considering the performance-efficiency trade-off, we establish grid size 6 with 16 calibration samples as the recommended default configuration. This setting achieves 71.3% average accuracy while consuming 0.9GB of memory and requiring 1020 seconds per layer, offering an optimal balance between quantization quality and computational overhead.

Table 3. Matrices needed to perform calibration and their sizes. C_o and C_i denote the output channel and input channel of weights, while b denotes the blocksize for lazy-batch update.

Matrix	GPTQ	GPTAQ	VLM-PTQ
W	$C_o \times C_i$	$C_o \times C_i$	$C_o \times C_i$
$\hat{W}$	$C_o \times C_i$	$C_o \times C_i$	$C_o \times C_i$
L	$C_i \times C_i$	$C_i \times C_i$	$C_i \times C_i$
P	-	$C_i \times C_i$	$C_i \times C_i$
C	-	-	C_i
M	-	-	C_i
W_b	$C_o \times b$	$C_o \times b$	$C_o \times b$
E_b	$C_o \times b$	$C_o \times b$	$C_o \times b$
$\hat{W}_b$	$C_o \times b$	$C_o \times b$	$C_o \times b$
L_b	$b \times b$	$b \times b$	$b \times b$
P_b	-	$b \times b$	$b \times b$

Table 4. Ablation study on grid search granularity and calibration sample size for modality-aware quantization on Qwen2.5-VL-7B-Instruct under W3A16 setting. 'Grid' denotes the number of candidate μ values searched, 'Samples' denotes the number of calibration samples used for evaluation. 'Mem' and 'Time' represent memory usage and time consumption for calibrating one layer, respectively.

Grid	Samples	Mem	Time	Avg
6	4	0.7GB	995s	70.8
6	16	0.9GB	1020s	71.3
6	64	1.4GB	1035s	71.5
6	128	2.1GB	1048s	71.5
11	4	0.7GB	1065s	71.0
11	16	0.9GB	1095s	71.6
11	64	1.4GB	1112s	71.8
11	128	2.1GB	1128s	71.9

References

[1] Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. Quarot: Outlier-free 4-bit inference in rotated llms. *Advances in Neural Information Processing Systems*, 37:100213–100240, 2024. 2

[2] Shiyao Li, Yingchun Hu, Xuefei Ning, Xihui Liu, Ke Hong, Xiaotao Jia, Xiuhong Li, Yaqi Yan, Pei Ran, Guohao Dai, et al. Mbq: Modality-balanced quantization for large vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4167–4177, 2025. 1

[3] Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of machine learning and systems*, 6:87–100, 2024. 1

[4] Yufei Xue, Yushi Huang, Jiawei Shao, and Jun Zhang. Vlmq: Efficient post-training quantization for large vision-language models via hessian augmentation. *arXiv preprint arXiv:2508.03351*, 2025. 1