CVPR 2026

CVPR JUNE 3-7, 2026 DENVER COLORADO

LS-ViT: Least-Squares Hessian Based Block Reconstruction for Low-Bit Post-Training Quantization of Vision Transformers

Hyunha Hwang, Xuan Truong Nguyen, Hyuk-Jae Lee

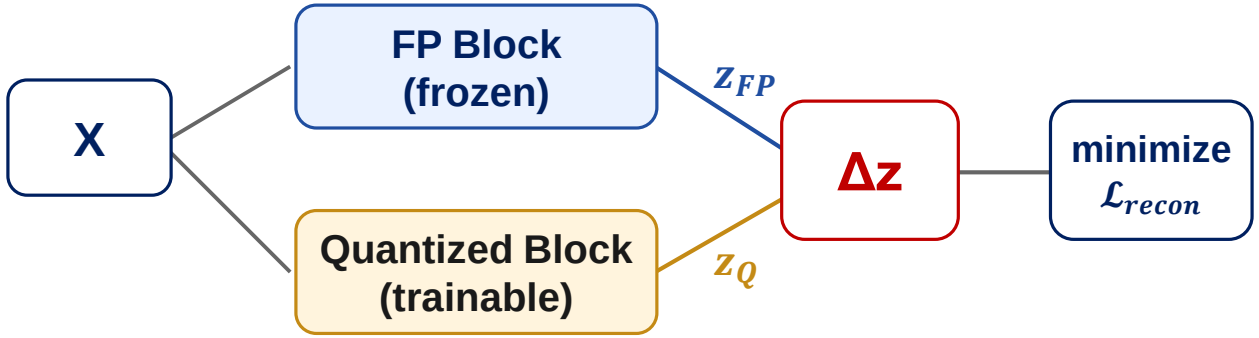
Presenter: Hyunha Hwang

Block Reconstruction and Hessian Metric

- Block Reconstruction Minimizes Quantization-Induced Loss
  - Block reconstruction matches the output of a quantized block to its full-precision counterpart.
  - Only quantization-related parameters are updated.
    - ✓ weight rounding & activation scales
  - The loss increase caused by quantization can be approximated using Taylor expansion.
  - Previous works ignore the first-order term and focus on the second-order Hessian term.

$$\mathcal{L}_{\text{recon}}(\Delta \boldsymbol{z}) = \frac{1}{2}\Delta \boldsymbol{z}^{\top}\mathbf{H}^{(\boldsymbol{z})}\Delta \boldsymbol{z}$$

  - Computing the full Hessian $\mathbf{H}^{(\boldsymbol{z})} = \nabla_{\boldsymbol{z}}^{2}\mathcal{L}$ is expensive.



Derivation

❖ Gradient–Perturbation Relation

- This paper assumes that $\mathbf{H}^{(z)}$ is symmetric and locally constant with respect to $\Delta \boldsymbol{z}$.
- By differentiating the reconstruction objective with respect to $\Delta \boldsymbol{z}$:

$$\nabla_{\Delta \boldsymbol{z}} \mathcal{L}_{\text{recon}}(\Delta \boldsymbol{z}) = \boldsymbol{g} = \mathbf{H}^{(z)} \Delta \boldsymbol{z}$$

- This equation connects three quantities:
  - ✓ gradient $\boldsymbol{g}$
  - ✓ block output perturbation $\Delta \boldsymbol{z}$
  - ✓ Hessian $\mathbf{H}^{(z)}$
- Hessian approximation can be viewed as fitting the relation between $\boldsymbol{g}$ and $\Delta \boldsymbol{z}$.

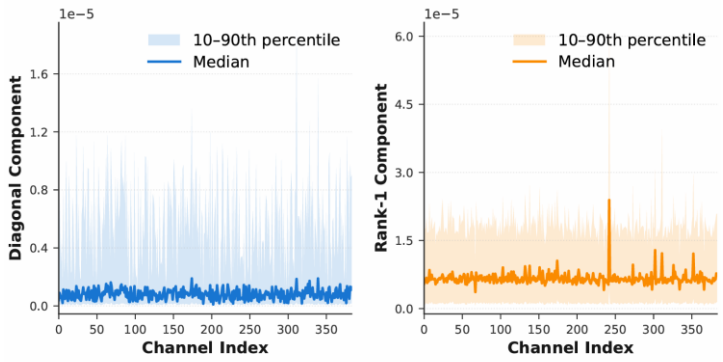
Limitation of Previous Methods

❖ Sample Independence Assumption

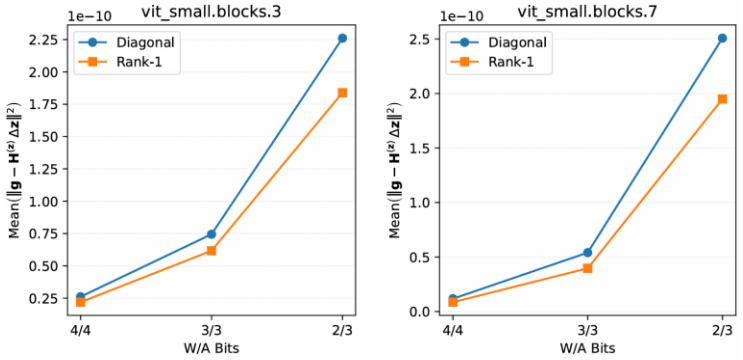
- Recent methods approximate the Hessian using averaged sample information.
  - ✓ FIMA-Q averages statistics across samples: $\mathbb{E}[\boldsymbol{g}] = \bar{\mathbf{H}}\mathbb{E}[\Delta \boldsymbol{z}]$.
  - ✓ These approaches assume independence between samples.
- However, this assumption ignores the covariance term:

$$\mathbb{E}\left[\mathbf{H}^{(\boldsymbol{z})}\Delta \boldsymbol{z}\right] = \mathbb{E}\left[\mathbf{H}^{(\boldsymbol{z})}\right]\mathbb{E}[\Delta \boldsymbol{z}] + \mathrm{Cov}(\mathbf{H}^{(\boldsymbol{z})}, \Delta \boldsymbol{z})$$

- This paper shows that the Hessian exhibits substantial sample-wise variation.
- As bit-width decreases, the covariance-induced approximation error becomes significant.



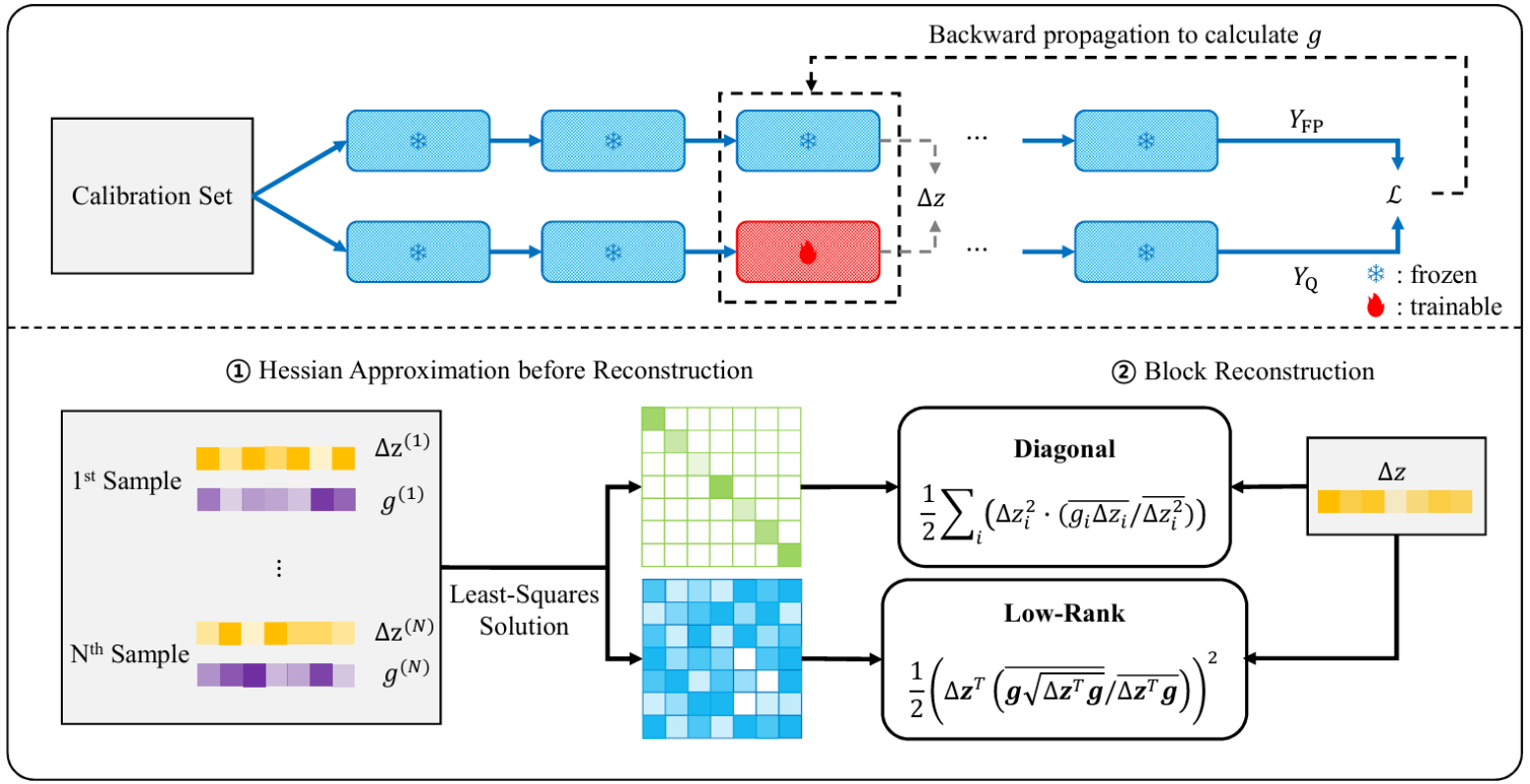
(a) Diagonal Component (b) Rank-1 Component



(a) vit_small.blocks.3 (b) vit_small.blocks.7

Proposed Method Pipeline

❖ LS-ViT: Two-Stage Reconstruction Process



- Stage 1: Calculate least-squares Hessian before reconstruction.
  - ✓ Use all pairs of ($\boldsymbol{g}$, $\Delta \boldsymbol{z}$) from the calibration set.
- Stage 2: Perform block reconstruction using the estimated Hessian.
  - ✓ Update weight roundings and activation scaling parameters.

Least-Squares Hessian Approximation

- Core Idea of LS-ViT

$$\widehat{\mathbf{H}} = argmin_{\mathbf{H}}\, \mathbb{E}[\left\| \mathbf{H}^{(z)}\Delta \boldsymbol{z} \; - \mathbf{H}\Delta \boldsymbol{z} \right\|^2$$

  - Instead of averaging sample-wise Hessians, LS-ViT estimates one shared Hessian via least-squares regression.
  - The estimated Hessian explicitly minimizes approximation error across samples.
  - This makes the Hessian more representative for block reconstruction.

Reconstruction Metric

❖ Least-Squares Hessian Approximation

- LS-ViT uses two complementary approximations

✓ Diagonal Hessian Approximation ($\mathbf{H}^{(\boldsymbol{z})} = \mathrm{diag}(H_{1,1}^{(\boldsymbol{z})},\ H_{2,2}^{(\boldsymbol{z})},\ \ldots, H_{d,d}^{(\boldsymbol{z})})$)

- $\widehat{H}_{i,i} = \frac{\sum_{n=1}^{N} g_i^{(n)} \Delta z_i^{(n)}}{\sum_{n=1}^{N} \left(\Delta z_i^{(n)}\right)^2} \rightarrow \mathcal{L}_{\mathrm{LSH,D}} = \frac{1}{2} \sum_i \Delta z_i^2 \left(\overline{g_i \Delta z_i} / \overline{(\Delta z_i)^2} \right)$
- captures individual sensitivity of each parameter.

✓ Rank-1 Hessian Approximation ($\mathbf{H}^{(\boldsymbol{z})} = \mathbf{u}\mathbf{u}^{\top}$)

- $\hat{\mathbf{u}} = \frac{\sum_{n=1}^{N} \boldsymbol{g}^{(n)} \sqrt{\Delta \boldsymbol{z}^{(n)\top} \boldsymbol{g}^{(n)}}}{\sum_{n=1}^{N} \Delta \boldsymbol{z}^{(n)\top} \boldsymbol{g}^{(n)}} \rightarrow \mathcal{L}_{\mathrm{LSH},L} = \frac{1}{2} \left(\Delta \boldsymbol{z}^{\top} \overline{\boldsymbol{g} \sqrt{\Delta \boldsymbol{z}^{\top} \boldsymbol{g}}} \,/\, \overline{\Delta \boldsymbol{z}^{\top} \boldsymbol{g}} \right)^2$
- captures dominant off-diagonal interactions.

✓ Final reconstruction loss combines both terms.

$$\mathcal{L}_{\mathrm{LSH}} = \mathcal{L}_{\mathrm{LSH,D}} + \mathcal{L}_{\mathrm{LSH},L}$$

Experimental Results

State-of-the-Art Accuracy

Method	W/A	ViT-S	ViT-B	DeiT-T	DeiT-S	DeiT-B	Swin-S	Swin-B
Full-Prec	32/32	81.39	84.54	72.71	79.85	81.80	83.23	85.27
PTQ4ViT [40]	3/3	0.10	0.10	3.50	0.10	31.06	28.69	20.13
RepQ-ViT [15]	3/3	0.10	0.10	0.10	0.10	0.10	0.10	0.10
GPTQ [7]	3/3	1.17	0.28	1.72	5.89	27.95	27.24	11.14
I&S-ViT [42]	3/3	45.16	63.77	41.52	55.78	73.30	74.20	69.30
DopQ-ViT [38]	3/3	54.72	65.76	44.71	59.26	74.91	74.77	69.63
QDrop* [32]	3/3	41.05	74.75	46.88	50.95	72.97	74.67	76.57
BRECQ* [13]	3/3	54.33	66.62	49.27	63.72	72.82	75.20	77.64
PD-Quant* [20]	3/3	54.73	73.13	53.79	67.52	75.93	76.62	78.66
APHQ-ViT [36]	3/3	63.17	76.31	55.42	68.76	76.31	76.10	78.14
APHQ-ViT(-)* [36]	3/3	59.11	76.05	53.82	67.40	75.89	75.44	77.31
FIMA-Q* [35]	3/3	64.09	77.63	55.55	69.13	76.54	77.26	78.82
LS-ViT (ours)	3/3	**64.10**	**77.65**	**55.72**	**69.41**	**76.57**	**77.39**	**79.40**
QDrop* [32]	2/4	49.35	71.44	38.87	45.99	70.01	70.13	68.64
BRECQ* [13]	2/4	45.20	67.18	46.31	61.20	69.72	69.21	70.36
PD-Quant* [20]	2/4	45.12	69.82	46.86	61.26	71.99	71.51	73.03
APHQ-ViT [36]	2/4	56.04	72.85	49.79	65.02	73.07	71.75	72.64
APHQ-ViT(-)* [36]	2/4	54.68	73.82	50.65	65.52	73.28	71.57	72.12
FIMA-Q* [35]	2/4	56.76	74.41	50.70	65.38	73.32	70.64	72.36
LS-ViT (ours)	2/4	**58.48**	**74.91**	**52.12**	**65.66**	**73.78**	**73.89**	**74.95**
QDrop* [32]	2/3	17.90	54.22	23.58	21.48	54.57	57.28	57.57
PD-Quant* [20]	2/3	25.56	46.06	31.45	48.66	63.60	62.41	65.29
APHQ-ViT [36]	2/3	41.63	58.06	38.33	53.53	66.09	62.90	65.03
APHQ-ViT(-)* [36]	2/3	36.78	58.00	37.12	52.78	66.02	59.83	61.55
FIMA-Q* [35]	2/3	38.82	61.15	36.93	52.57	66.24	59.57	63.15
LS-ViT (ours)	2/3	**42.05**	**62.89**	**39.28**	**53.86**	**67.06**	**65.77**	**67.73**

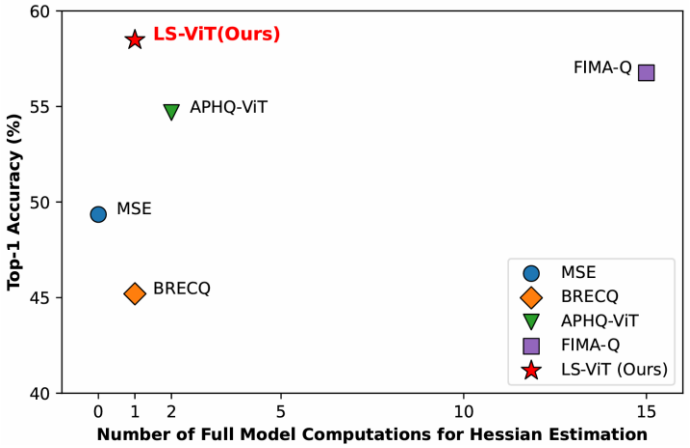
Method	W/A	Mask R-CNN				Cascade Mask R-CNN			
		Swin-T		Swin-S		Swin-T		Swin-S	
		AP^b	AP^m	AP^b	AP^m	AP^b	AP^m	AP^b	AP^m
Full-Precision	32/32	46.0	41.6	48.5	43.3	50.4	43.7	51.9	45.0
Baseline	4/4	34.6	34.2	40.8	38.6	45.9	40.2	47.9	41.6
PTQ4ViT [40]	4/4	6.9	7.0	26.7	26.6	14.7	13.5	0.5	0.5
APQ-ViT [4]	4/4	23.7	22.6	**44.7**	40.1	27.2	24.4	47.7	41.1
RepQ-ViT [15]	4/4	36.1	36.0	44.2	40.2	47.0	41.1	49.3	43.1
GPTQ [7]	4/4	36.3	36.3	42.9	40.2	47.1	41.5	49.2	43.2
ERQ [41]	4/4	36.8	36.6	43.4	40.7	47.9	42.1	50.0	43.6
I&S-ViT [42]	4/4	37.5	36.6	43.4	40.3	48.2	42.0	50.3	43.6
DopQ-ViT [38]	4/4	37.5	36.5	43.5	40.4	48.2	42.1	50.3	43.7
QDrop* [32]	4/4	36.2	35.4	41.6	39.2	47.0	41.3	49.0	42.5
FIMA-Q* [35]	4/4	38.7	37.8	44.2	**41.1**	48.7	42.5	50.4	43.7
LS-ViT (ours)	4/4	**38.9**	**38.0**	44.3	**41.1**	**48.8**	**42.6**	**50.6**	**43.9**

Efficiency

- High Accuracy with Lower Quantization Cost
  - LS-ViT requires only one full-model computation per block.
  - It is 1.8× to 2.7× faster than FIMA-Q.
  - Training time is close to QDrop while achieving better accuracy.

Table 5. Training time (minutes) comparison for each method using a single GPU.

Model	QDrop	FIMA-Q	LS-ViT (ours)	Improvement
DeiT-T	100	180	**100**	**1.8**×
DeiT-S	105	225	**105**	**2.1**×
DeiT-B	145	310	**150**	**2.1**×
Swin-S	160	420	**165**	**2.5**×
Swin-B	170	480	**180**	**2.7**×



Ablation Study

❖ Impact of Independence and Averaging

- Problem with sample-wise Hessian
  - ✓ Numerically unstable when Δz_i is small
  - ✓ High variance across calibration samples
- Limitation of simple averaging
  - ✓ Reduces variance
  - ✓ But loses sample-wise relationship between g_i and Δz_i
- Least-squares aggregation reduces variance while preserving sample relationships
  - ✓ FIMA-Q-D is a simplified case of our formula under the assumption of sample independence

$$\widehat{H}_{i,i} = \frac{\mathbb{E}\left[g_i \Delta z_i\right]}{\mathbb{E}\left[\Delta z_i^2\right]} = \frac{\mathrm{Cov}(g_i, \Delta z_i) + \mathbb{E}[g_i]\mathbb{E}[\Delta z_i]}{\mathrm{Var}(\Delta z_i) + (\mathbb{E}[\Delta z_i])^2}$$

Method	Approx.	W/A	ViT-S	ViT-B	DeiT-T	DeiT-S	DeiT-B	Swin-S	Swin-B
Sample-Wise	$g_i/\Delta z_i$	3/3	23.59	46.22	37.92	50.50	66.42	69.42	73.25
Sample-Wise	g_i	3/3	58.16	74.34	54.64	68.03	75.61	76.02	78.22
FIMA-Q-D	$\mathbb{E}\left[g_i\right]/\mathbb{E}\left[\Delta z_i\right]$	3/3	60.02	76.29	55.54	68.68	76.32	75.08	77.87
LS-ViT-D (ours)	$\mathbb{E}\left[g_i \Delta z_i\right]/\mathbb{E}\left[\Delta z_i^2\right]$	3/3	**63.25**	**77.35**	**55.55**	**69.30**	**76.55**	**77.29**	**79.03**

Conclusion

- LS-ViT provides accurate and efficient low-bit PTQ for practical ViT deployment.
  - LS-ViT formulates Hessian approximation as a regression problem over calibration samples.
  - It captures sample-wise relationships ignored by averaging-based methods.
  - It achieves state-of-the-art low-bit PTQ accuracy.
  - It reduces quantization cost by requiring only one full-model computation per block.