Comp²VLM

A Hybrid Framework Combining Quantization and Lossless Compression for Efficient Vision-Language Models

Dahun Choi*, **Inseong Hwang***, **Hyun Kim†** *Equal contribution †Corresponding author

Department of Electrical and Information Engineering, Seoul National University of Science and Technology, Seoul, Korea · {dahun926, ish9707, hyunkim}@seoultech.ac.kr

NeurIPS 2026

40th Conference on Neural Information Processing Systems

W3.5A4.6
effective bits on LLaVA-OV-7B, beating W4A8 MBQ

−33%
entropy of quantized weights with DEQ

~4%
end-to-end latency overhead of QPress (Huffman 65–72%)

Why combine the two?

VLMs keep scaling: Qwen2-VL-72B needs about 146 GB just for inference. **Quantization** shrinks it, but low-bit activations break the fine-grained visual–linguistic alignment VLMs depend on. **Lossless coding** keeps accuracy exactly, but its variable-length streams and heavy metadata fight GPU parallelism.

Naively stacking them at low bit-width hits two problems.

C1 Accuracy vs. compressibility

Standard PTQ spreads values over as many bins as possible to minimize error. That maximizes entropy and leaves nothing for the entropy coder to exploit.

DEQ picks per-group scales that trade a little MSE for many more repeated symbols.

C2 Metadata and decoding cost

Low-bit data is already compact, so per-chunk codebooks dominate storage, and sequential Huffman decoding stalls the accelerator.

QPress shares 64 universal Huffman tables and decodes 256-byte blocks independently.

Prior double compression

	WCM	QStore	Ours
Validated below 8-bit	×	×	✓
Evaluated on VLMs	×	×	✓
Shared metadata	—	×	✓
Parallel decoding	—	×	✓

Experimental setup

Qwen2-VL-7B/72B, LLaVA-OV-7B/72B and InternVL2-8B on TextVQA, OCRBench, ScienceQA, VizWiz and SEED (LMMs-Eval, one H100). τ and the Huffman tables are fixed at calibration on a 5% random subset of the ScienceQA, TextVQA and VizWiz training splits (3,392 samples, no evaluation instances); one table bank per model and precision is shared by every benchmark.

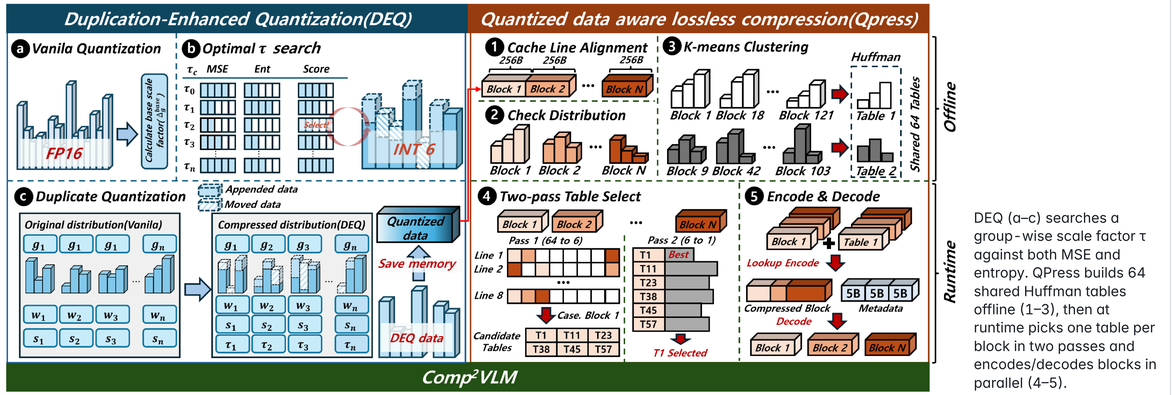
Does calibration data leak into the rate?

Calibration source	SciQA A	TextVQA A	W
Default subset	3.42	3.20	3.53
COCO Captions (disjoint)	**3.43**	**3.21**	3.53
VizWiz only	3.45	3.23	3.53

Effective bits, Qwen2-VL-7B (W6A6). A fully disjoint source costs only 0.01 bits.

Scope

Built for the 5–7-bit regime, where quantization alone still loses accuracy and entropy coding still finds redundancy. Default is W6A6, which packs and vectorizes better than W4A8. Not designed for sub-4-bit.



DEQ (a–c) searches a group-wise scale factor τ against both MSE and entropy. QPress builds 64 shared Huffman tables offline (1–3), then at runtime picks one table per block in two passes and encodes/decodes blocks in parallel (4–5).

DEQ Duplication-Enhanced Quantization

Enlarging the step size collapses more values into the same integer bin: lower entropy, slightly higher error. DEQ scores every candidate scale on both and keeps the best per group.

$$\Delta_g(\tau_c) = \tau_c \cdot \max_j |X_{g,j}| \;/\; q_{\max} \qquad \text{candidate scale}$$

$$S_g(\tau_c) = \alpha\, \widetilde{D}_g(\tau_c) + (1-\alpha)\, \widetilde{H}_g(\tau_c) \qquad \text{score}$$

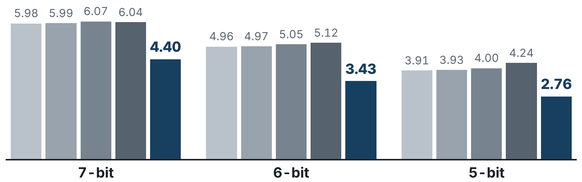
$$\tau_g^{\star} = \arg\min_{\tau_c \in \mathscr{T}} S_g(\tau_c) \qquad \alpha = 0.4$$

$\widetilde{D}$: group MSE, $\widetilde{H}$: Shannon entropy of the integer codes, both min–max normalized over $\mathscr{T}$ so groups of different magnitude are comparable. α = 1 reduces to RTN.

Search per group

1. Quantize the group with every τ_c in the candidate set $\mathscr{T}$.
2. Measure MSE and the entropy of the resulting integer codes.
3. Normalize both over $\mathscr{T}$ and keep the τ with the lowest score.

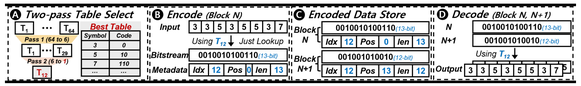
Entropy of quantized weights (bits, InternVL2-8B)



26.4–33.0% lower than every baseline at each precision. Since entropy lower-bounds lossless code length, the ~1.4-bit average drop below RTN leaves room for about 28% more compression at the same nominal bit-width.

QPress Quantized data-aware lossless compression

Low-bit codes have a tiny alphabet, so a few shared tables cover nearly every block. Tables come from K-means over symbol histograms, built once per model.



Each block stores only table index, bit offset and length, so any thread can locate and decode its block by lookup.

64
shared Huffman tables per bank (W, A)

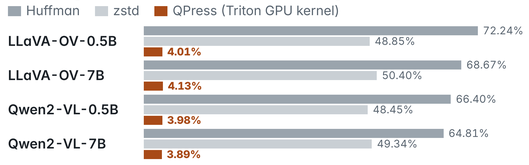
5 B
metadata per 256-byte aligned block

9.4%
of greedy table-search cost (64 → 6 → 1)

Compression gain over baselines (Qwen2-VL-7B)

	vs. Huffman		vs. zstd	
Precision	**W**	**A**	**W**	**A**
5-bit	8.2%	**13.3%**	11.5%	3.5%
6-bit	7.5%	12.2%	8.2%	2.2%
7-bit	5.9%	10.7%	5.4%	1.7%
Average	**7.2%**	**12.1%**	**8.4%**	**2.5%**

Latency overhead over quant-only inference



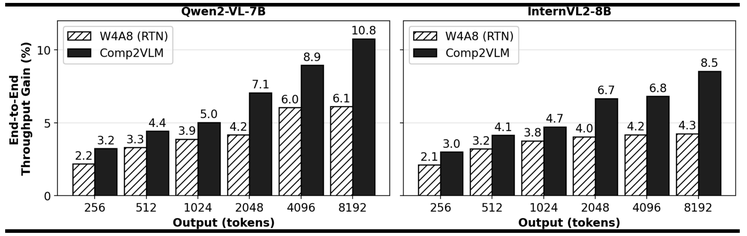
Results DEQ + QPress from a W6A6 start

Accuracy (%) vs. W4A8 baselines

Model	Method	SciQA	TextVQA	VizWiz	OCRB.	SEED
Qwen2-VL-7B W3.5 A3.2–3.4	BF16	84.7	84.4	68.8	84.2	76.4
	MBQ	81.2	75.0	59.3	72.8	75.1
	MQuant	–	84.3	–	82.4	–
	Ours	**84.1**	**83.3**	**66.8**	**81.0**	**74.8**
LLaVA-OV-7B W3.5 A4.5–4.9	BF16	90.0	76.1	60.4	62.2	72.0
	MBQ	88.5	68.3	58.9	52.3	67.7
	Ours	**89.3**	**73.0**	**58.9**	**60.0**	**71.3**
InternVL2-8B W3.6 A4.6–4.7	BF16	96.2	77.7	60.7	76.5	71.6
	MBQ	95.8	72.3	56.5	73.0	67.7
	MQuant	–	76.6	–	72.5	–
	Ours	**95.9**	**75.5**	**61.9**	**72.7**	**71.2**
Qwen2-VL-72B W3.5 A3.5–3.6	BF16	92.5	85.5	75.8	88.3	77.6
	MBQ	89.6	80.5	73.6	77.5	76.3
	MQuant	–	85.0	–	86.8	–
	Ours	**92.1**	**85.5**	**75.2**	**87.7**	**76.5**
LLaVA-OV-72B W3.5 A4.6–5.0	BF16	90.0	79.3	69.2	73.2	78.1
	MBQ	89.2	73.3	65.7	64.4	76.5
	Ours	**89.9**	**77.4**	**66.2**	**70.1**	**78.0**

MBQ and MQuant use W4A8. Green labels give Comp²VLM's effective bit-width after DEQ and QPress; activation bits vary with input entropy.

End-to-end throughput gain over W4A8 RTN (H100, batch 1)



Same Triton low-bit kernel family for both. The gap widens with output length (up to 10.8% / 8.5% at 8192 tokens) because decoding is memory-bound and W3.5 cuts weight traffic by ~12.5% vs. W4.

On-device: Jetson AGX Orin (Qwen2-VL-7B)

Method	Tok/s	J/token	Peak memory
BF16	12.41	3.44	13.64 GB
RTN W6A6	26.26	1.88	5.59 GB
Ours	**26.24**	**1.87**	**3.42 GB**

Same throughput as plain W6A6 (−0.08%) with 39% less peak memory; bit-exact reconstruction verified on-device.

[MBQ] Li et al., CVPR 2025 · [MQuant] Yu et al., ACM MM 2025 · [QuaRot] Ashkboos et al., NeurIPS 2024 · [SQ] Xiao et al., ICML 2023 · [WCM] Wang et al., EMNLP Findings 2024 · [QStore] Shah et al., arXiv 2025 · [DFloat11] Zhang et al., arXiv 2025

Supported by NRF (RS-2026-25481790) and MOTIE Technology Innovation Program (RS-2025-02307330)