Less Precise Can Be More Reliable:

A Systematic Evaluation of Quantization's Impact on VLMs Beyond Accuracy.

Aymen Bouguerra, Daniel Montoya, Alexandra Gomez-Villa, Chokri Mraidha, and Fabio Arnez.

ICML International Conference On Machine Learning 2026 SEOUL · 6–11 JUL

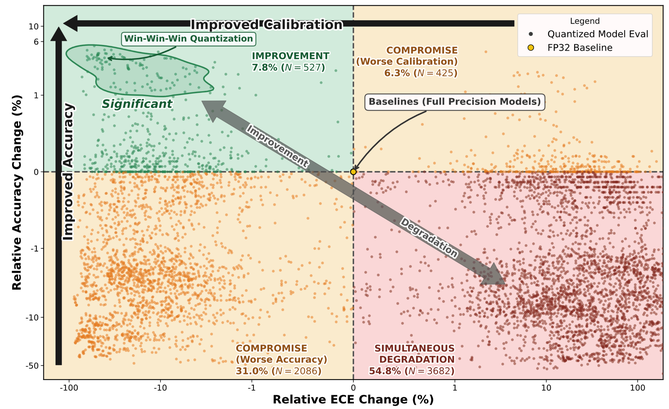
Quantization can make models more reliable: Can Improve : Accuracy · Calibration · OOD Detection · Corruption Robustness | Can Degrade : Semantic Robustness · Spurious Correlation Robustness

8k+ Quantized Models | **700k+ Configurations Tested** | **Zero-shot Domain** | **PTQ & QAT Quantization** | **Vision-only & Vision + Text Quantization**

PROBLEM & GAP

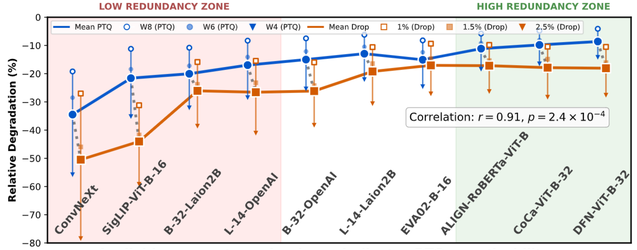
- VLMs are the default for zero-shot and safety-critical OOD.
- Deployment forces quantization, and it is evaluated solely on Top-1 accuracy.
- Reliability under quantization is an unmeasured blind spot.

IMPACT



Less precise, more reliable.

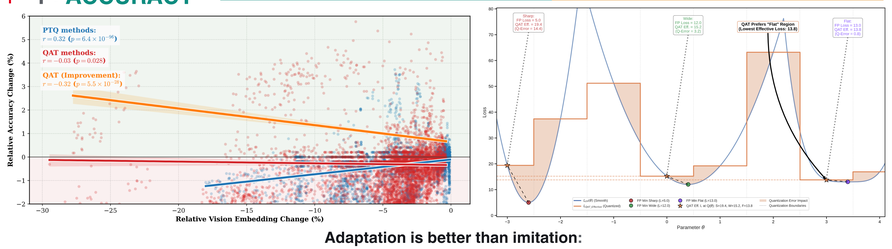
~40% of W8A8 runs improve calibration; 8.5% improve BOTH accuracy and calibration vs FP32.



Available parameter redundancy indicates PTQ quantizability:

A strong correlation between degradation under PTQ (blue) and weight dropout (orange) indicates that redundancy is the primary robustness buffer.

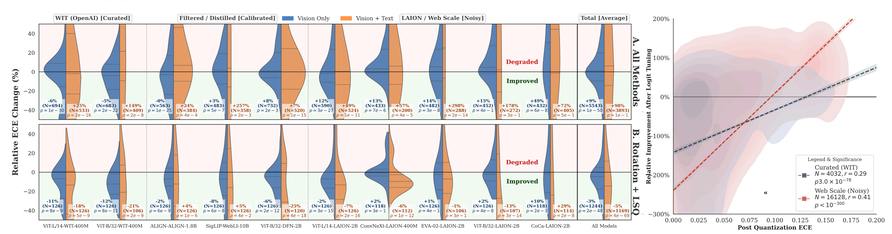
1 · ACCURACY



Adaptation is better than imitation:

Embedding similarity correlates with good PTQ quantization. However, good quantization adaptation (QAT) and improvement correlate with embedding dissimilarity.

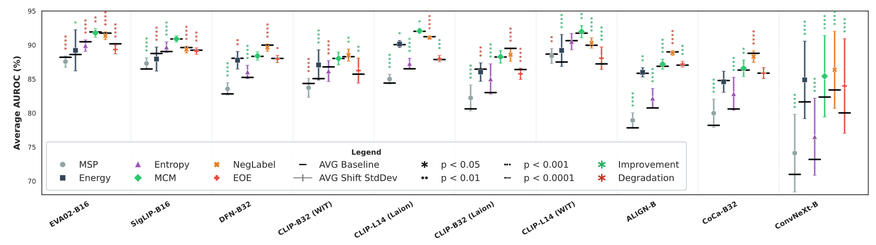
2 · CALIBRATION



Calibration is driven by initial calibration:

ECE improves overall, but Quantization's impact depends heavily on the initial calibration of the model; a mis-calibrated model tends to improve, while a calibrated one tends to worsen

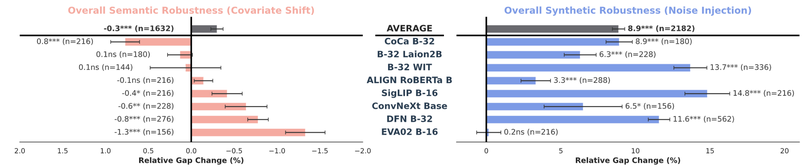
3 · OOD DETECTION



Accuracy & AUROC move together:

Tightly coupled (r = 0.994, N = 466k); VLM-specific scorers (MCM,NegLabel) stay strong under W8A8.

4 · COVARIATE SHIFT

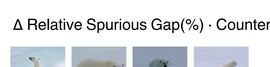


The shift dichotomy:

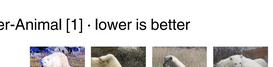
Quantization improves the model's robustness to corruption but degrades its robustness to semantic and stylistic shift.

5 · SPURIOUS BIAS

Δ Relative Spurious Gap(%) · Counter-Animal [1] · lower is better



Photos of **ice bear** in **snow** background (**easy**, accu 97.62)



Photos of **ice bear** in **grass** background (**hard**, accu 70.91)

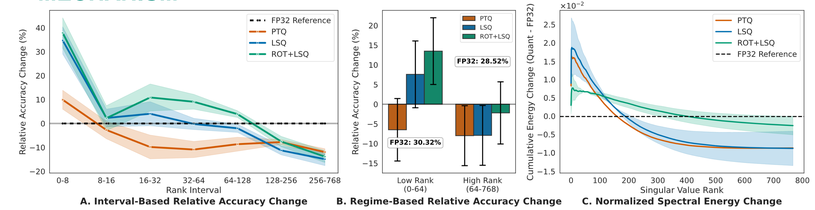
[1] Q. Wang et al. A Sober Look at the Robustness of CLIPs to Spurious Features. NeurIPS 2024.

METHOD	W8A8	W6A8	W4A8
Ordinary PTQ	+2.6	+2.8	+12.5
QAT (Contrastive)	+1.6	+2.1	+9.1
Rotation + LSQ	−0.1	+0.2	+4.0

Quantization can only degrade spurious correlation robustness:

Even when quantization improves accuracy, it still causes the model to be more reliant on spurious features.

MECHANISM



Quantization amplifies spectral bias:

A - Where does the quantized model improve? And where does it degrade?
We predict using both the original model and its quantized version using only intervals of top-k SVD ranks; lower ranks tend to represent general features while higher ranks represent specialist features.

B - How does quantization impact general (low rank) features vs specialist (high rank) features?
We aggregate the results from plot A and separate the comparison into two regimes: low rank and high rank.

C - Quantization-induced energy shift.
Quantized models have less energy in the high-rank tail while having more in the lower ranks.

DEPLOYMENT & PRACTICE: FASTER, SMALLER & MORE Reliable

CLIP ViT-B/32 (WIT) · CIFAR-100 · 10 seeds · RTX 4070

ENGINE	ACCURACY (%)	LATENCY (ms)	SPEEDUP	SIZE (mb)
FP32 PyTorch	64.47	120.9	X1	302.1
FP16 TensorRT	64.43	22.7	X5.33	187.7
INT8 TensorRT	66.30	15.7	X7.72	118.4

Accuracy drop under Gaussian blur (pp, lower is better)

BLUR STRENGTH	FP16 Acc Drop(pp)	INT8 Acc Drop(pp)	ADDED ROBUSTNESS(pp)
Mild	3.92	1.71	+2.21
Moderate	12.28	7.99	+4.29
Severe	35.64	30.72	+4.92

TAKEAWAY

- **First large-scale reliability benchmark for quantized VLMs:** Quantization isn't uniformly destructive: it can strengthen calibration, OOD detection, and noise robustness but weakens shift robustness and worsens spurious bias.
- **One mechanism explains the split:** spectral low-pass filtering. Rounding noise buries high-rank, low-variance features and forces the model to rely more on coarse, robust ones; Quantization amplifies the spectral bias of VLMs.

How Does Quantization Impact Contrastive Visual-Language Models' Reliability ?

CORRESPONDENCE
aymen.bouguerra@cea.fr

université PARIS-SACLAY · CVC Computer Vision Center