CAR-SAM: Cross-Attention Reconstruction for Post-Training Quantization of the Segment Anything Model

Houji Wen[1,2], Jiangyong Yu[2]*, Dawei Yang[2], Jun Li[1]

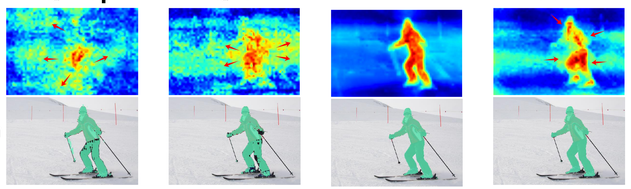
[1]Nanjing University of Science and Technology [2]Houmo AI

CVPR JUNE 3-7, 2026 DENVER COLORADO

Motivation

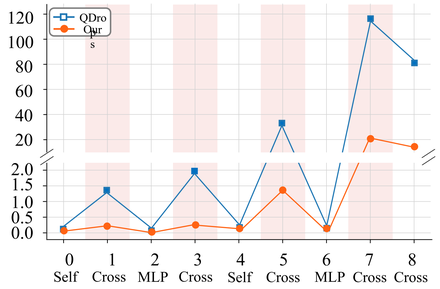
SAMs achieve strong prompt-guided segmentation, but their large memory and compute cost hinder deployment on edge devices. Existing PTQ methods mainly target encoder-only ViTs and do not explicitly handle the bidirectional cross-attention structure in the SAM decoder.

Attention dissipation



***Attention Dissipation* leads to degradation masks**

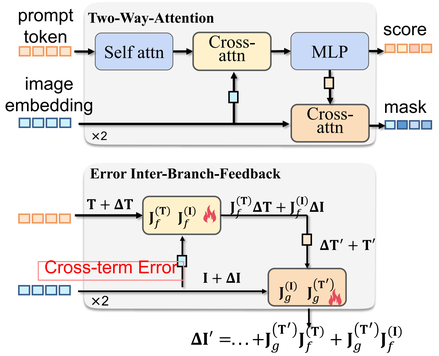
Reconstruction Loss Oscillation



Cross-attention activations fluctuate across decoder blocks, causing unstable reconstruction under low-bit quantization

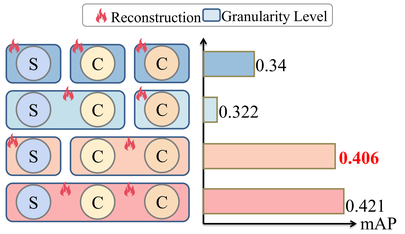
Method Overview

A) Jont Reconstruction



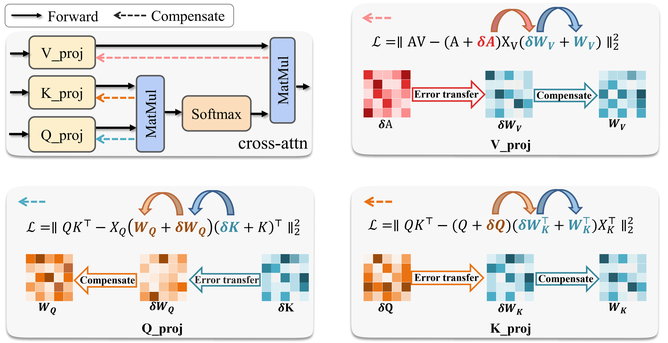
Joint reconstruct coupled cross-attention modules to suppress oscillation

Why Jont Reconstruction?



Joint reconstruction of the two cross-atten modules yields the highest accuracy.

B) MatMul-Aware Compensation



Compensate Matmul-Induced activation error by updating Q/K/V projection weight

1) Objections: Minimize MatMul Error

$$\mathcal{L} = \| QK^\top - X_Q(W_Q + \delta W_Q)(\delta K + K)^\top \|_2^2$$

2) Sylvester Reformulatiuon

$$AX + XB = C$$

where

$$\mathbf{A} = \left(\mathbf{X}_Q^\top \mathbf{X}_Q\right)^{-1} \lambda \mathbf{I}$$
$$\mathbf{B} = \widehat{\mathbf{K}}^\top \widehat{\mathbf{K}}$$
$$\mathbf{C} = \mathbf{W}_Q \delta \mathbf{K}^\top \widehat{\mathbf{K}}.$$
$$\mathbf{X} = \delta \mathbf{W}_Q$$

3) Weight update

$$\mathbf{W}_Q = \mathbf{W}_Q + \delta \mathbf{W}_Q$$
$$\mathbf{W}_K = \mathbf{W}_K + \delta \mathbf{W}_K$$
$$\mathbf{W}_V = \mathbf{W}_V + \delta \mathbf{W}_V$$

Conclusion

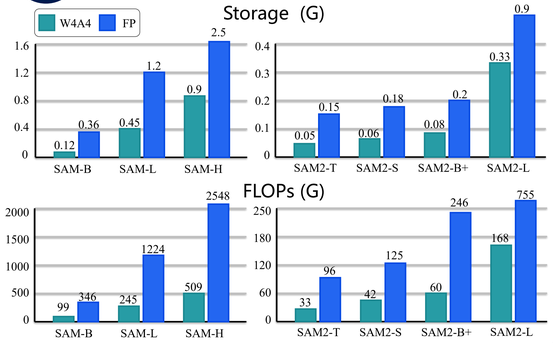
CAR-SAM identifies two decoder-specific quantization failures in SAM: attention dissipation and reconstruction oscillation. MAC compensates MatMul-induced errors, while JCAR stabilizes coupled cross-attention reconstruction, enabling robust 4-bit quantization for SAM and SAM2.

Main result (COCO W4A4)

Model	QDrop	PTQ4SAM	CAR-SAM (Ours)
SAM-B	26.4	24.7	**39.3**
SAM-L	38.0	41.9	**48.5**
SAM-H	48.8	50.7	**51.6**

+14.6% mAP on SAM-B and +6.6% on SAM-L over prior methods at 4bit

Efficiency (W4A4 vs FP16)



Qualitative Result