WHERE PRECISION MATTERS: SELECTION-REGRET LEARNING FOR QUANTIZED SAM 2

Anonymous authors
Paper under double-blind review

ABSTRACT

Segment Anything Model 2 (SAM 2) generates multiple mask candidates for each prompt and selects one using its predicted-IoU head, but existing quantization objectives evaluate candidates independently and do not protect this selection process. We propose a mixed-precision quantization method that jointly learns the weights and a bit width for every layer, together with a teacher-relative selection-regret objective that penalizes only the additional mask quality forfeited by the quantized selector beyond its full-precision teacher. A controlled precision sweep over video propagation units shows that hard selection accuracy degrades significantly already at eight bits, even when candidate quality remains within half a percent of full precision, and that the damage increases with propagation depth. Precision learned using only the segmentation loss concentrates around the same predicted-IoU interface, which two independent routes single out. At an average precision of 4.79 bits, our method achieves 71.1 SA-V $\mathcal{J}\&\mathcal{F}$ and 0.743 COCO mIoU on SAM 2.1 Base+, whereas two mixed-precision baselines at the same budget preserve COCO accuracy but fall below 2.4 on video. At a matched budget, the regret objective raises candidate-set and selected-mask quality, each with a paired interval excluding zero.

1 INTRODUCTION

Segment Anything Model 2 (SAM 2) (Ravi et al., 2025) segments images and videos from simple prompts without task-specific training. Widely used for annotation and tracking, its Base+ variant has 80.9M parameters at 323 MB and processes every video frame with memory. Quantization reduces this cost (Han et al., 2015; Lin et al., 2016; Cheng et al., 2017; Jacob et al., 2018; Banner et al., 2019; Choukroun et al., 2019; Nagel et al., 2021), while mixed precision assigns different bit widths according to layer sensitivity (Dong et al., 2020; Wang et al., 2019; Yang & Jin, 2021). Below five bits the weights fit in under 50 MB, but identifying precision-critical layers remains difficult. SAM and SAM 2 methods use static calibration or fixed assignments (Lv et al., 2024; Liu et al., 2024; Ranjan & Savakis, 2025; Farronato et al., 2025; Duan et al., 2026) without explaining what makes a layer precision-critical.

SAM 2 degrades sharply: six-bit execution remains usable, but four-bit execution collapses at two of three scales. Its predicted-IoU head scores multiple mask candidates and returns the highest-scoring one. In video, that mask also enters memory and conditions later frames, coupling candidate generation and selection to error propagation.

Executing one fixed checkpoint at eight, six, five, and four bits shows that selection degrades before the masks do. At eight bits, selection accuracy falls significantly while candidate quality remains within half a percent of full precision (Figure 1a). At six bits, the hit rate drops a quarter against a 2.3% candidate-quality loss. At five bits, candidate quality falls 21% and selection errors cost six times as much. One slot's forfeited quality rises from 3.1% to 15.8%, the other two change by less than twofold, and their ordering changes (Figure 1b). Within trajectories, candidate separation grows with propagation depth while the full-precision control is flat (Figure 1c). These observations point at the predicted-IoU head, and testing the claim needs a learned allocation.

In this paper, we jointly learn layerwise bit widths and model weights, and supervise candidate selection with a teacher-relative selection-regret objective. Continuous bit-width parameters control

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

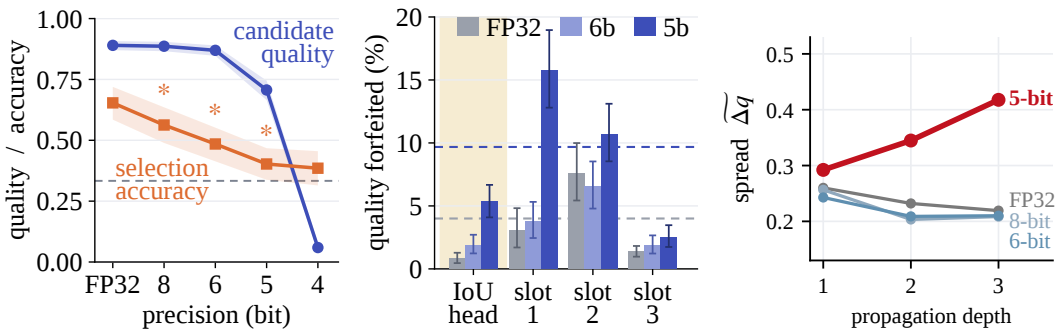

Figure 1: Selection degrades before the masks do. One fixed checkpoint at each precision on SA-V propagation frames, with bootstrap intervals. (a, left) Hard selection accuracy and oracle candidate quality, chance $1/3$, with asterisks on the accuracy curve where the paired drop from full precision excludes zero. (b, middle) Quality forfeited by the head and by each fixed slot at full precision, six bits, and five bits, with the uninformed policy dashed, least by the head at full precision and more than the best slot at five bits. (c, right) Candidate-quality separation, measured by the scale-free spread $\widetilde{\Delta q}$, against propagation depth on SA-V. As shown in the graph, the five-bit curve rises with depth, and the others do not.

forward-pass quantization step sizes and are optimized with the weights under the segmentation loss. Even without a candidate-comparison term, this loss yields a profile whose highest retained precision lies in the predicted-IoU head (Figure 3), independently identifying the interface found by the controlled sweep.

This objective compares candidates rather than individual outputs. We detach candidate qualities to prevent direct mask optimization through the quality measure, and define regret as the quality forfeited relative to the best available candidate. Because the teacher can also select imperfectly, we penalize only the additional regret from quantization. Gradients pass through the predicted-IoU pathway and shared decoder and quantization parameters, potentially changing both the selection and candidate set. Our contributions are threefold.

Selection degrades before the masks do. A controlled precision sweep over video propagation units puts hard selection accuracy significantly below full precision already at eight bits, with candidate quality still within half a percent, and the damage grows with propagation depth and falls on one decoder slot rather than spreading evenly.

Two independent routes single out the same interface. The controlled sweep and a profile learned under the segmentation loss alone converge on the predicted-IoU readout, by routes that share no measurement.

A selection-regret objective defined on the comparison rather than on a single output. At a matched budget the term raises the candidate mean and the selected-mask quality by $+0.05$, each with a paired interval excluding zero.

2 Related Work

Quantization for SAM and SAM 2. Most prior work targets SAM with post-training quantization, where careful calibration supports 8-bit operation and later refinements push precision lower with activation clipping and sensitivity analysis (Lv et al., 2024; Liu et al., 2024; Zhang et al., 2025; 2026), while a smaller line applies quantization-aware training to SAM-like models (Lu et al., 2025). SAM 2 has received less attention, and recent studies use improved activation handling, post-training calibration, or mixed-precision allocation (Farronato et al., 2025; Ren et al., 2025; Duan et al., 2026). All report the accuracy following an allocation without saying what lost precision disturbs, and we ask instead what low precision does to the candidate set the decoder chooses among at every propagated frame.

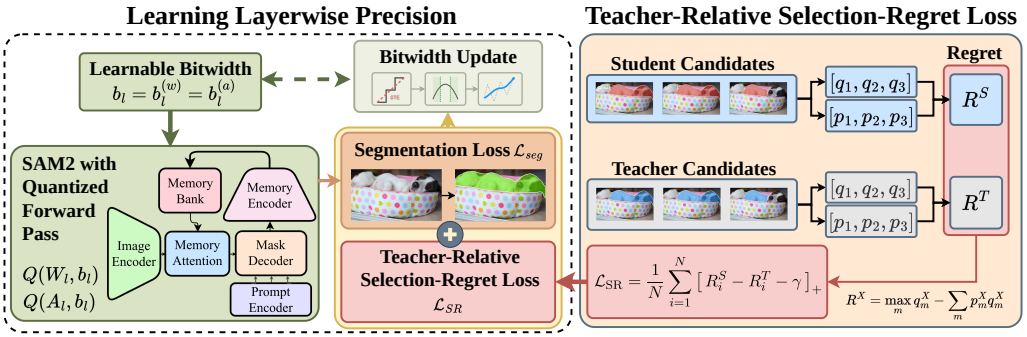

Figure 2: Overview of our proposed method. Left: a precision parameter per layer enters the quantized forward pass through the step size and is updated by the segmentation loss alongside the weights. Right: the student and a frozen full-precision teacher each score their own candidates, and $\mathcal{L}_{\mathrm{SR}}$ penalizes only the regret the student incurs beyond the teacher.

Mixed-precision quantization. Early methods estimate layer sensitivity with second-order approximations (Dong et al., 2019; 2020; Yao et al., 2021), while hardware-aware formulations fold device constraints into a search or a learned policy (Wang et al., 2019; Uhlich et al., 2020). Later work relaxes the discrete bit choice so that it can be optimized by gradient descent (Wu et al., 2018; Gong et al., 2019; Yang & Jin, 2021; Louizos et al., 2019), treats bit widths as parameters learned or pruned during training (Bhalgat et al., 2020; Stock et al., 2019; Nikolić et al., 2024), derives an allocation from a learned notion of layer importance (Tang et al., 2022), or applies the idea to SAM-style models (Ranjan & Savakis, 2025). In each case the allocation is the deliverable. We learn a profile the same way, relaxing the bit width and training under the task loss, but then use the profile as evidence rather than as the result, comparing the resulting interface against the interface found by a controlled precision sweep.

Behavior-preserving objectives. Quantization-aware training is often paired with an objective that preserves the behaviour of a full-precision teacher rather than the task label alone. Distillation transfers the teacher's soft outputs (Hinton et al., 2015), an idea adapted to low-precision students (Mishra & Marr, 2018; Polino et al., 2018; Kim et al., 2019) and extended to intermediate activations (Heo et al., 2019). All of these preserve behaviour defined pointwise, over a single output or feature map, whereas the decision SAM 2 makes at each propagated frame is a comparison among candidates, at a cost set by how far apart the candidates lie. Work on Dice- and IoU-calibrated segmentation makes the same point one level down, showing that an argmax or a threshold over per-pixel scores is inconsistent for those metrics and that the decision rule has to be derived from the metric itself (Wang & Dai, 2025).

3 Methodology

Figure 2 combines task-loss-driven layerwise precision learning with teacher-relative selection regret. Continuous precision parameters enter through quantization step sizes, while detached candidate qualities restrict the regret gradient to the predicted-IoU pathway.

3.1 Preliminaries

3.1.1 Quantization basics.

For a bit width b and a clipping range $[-\alpha, \alpha]$, uniform symmetric quantization has step size $\Delta(b) = \alpha/(2^{b-1}-1)$ and fake quantization operator

$$Q(x; b) = \Delta(b) \cdot \mathrm{clip}\Big(\Big\lfloor \tfrac{x}{\Delta(b)} \Big\rceil,\ -(2^{b-1}-1),\ 2^{b-1}-1\Big). \tag{1}$$

The rounding uses straight-through estimation during quantization-aware training and the clipping the identity gradient within the valid range. The weights stay in floating point and are quantized only in the forward pass, so a trained checkpoint can be executed at any width, including none.

3.1.2 BRIEF RECAP OF SAM 2.

Standard SAM 2 combines a hierarchical encoder, multi-scale neck, prompt encoder, mask decoder, and temporal memory. For one prompt, a shared decoder produces M mask tokens and one IoU token. The former yield M candidate masks, while a three-layer predicted-IoU head maps the latter to M scores $\mathbf{s} \in \mathbb{R}^M$. The highest-scoring mask is returned and stored in video memory. Predicted-overlap scoring is established (Goldman et al., 2019), while SAM 2 reads scores from one token and propagates the winner.

3.2 LEARNING LAYERWISE PRECISION

For each quantized layer l we associate a continuous precision parameter b_l, entering the forward computation through the step size $\Delta(b_l)$. That step size is differentiable in b_l,

$$\frac{\partial \Delta(b_l)}{\partial b_l} = -\Delta(b_l)\, \ln 2\, \frac{2^{b_l - 1}}{2^{b_l - 1} - 1}, \tag{2}$$

and strictly decreasing, so raising precision refines the grid and a gradient can reach b_l through Δ. The quantized weights and activations of layer l are $W_l^{(q)} = Q(W_l; b_l^{(w)})$ and $A_l^{(q)} = Q(A_l; b_l^{(a)})$. For activation quantization, the clipping threshold is learned jointly with the network parameters (Choi et al., 2018). In the simplest form, a layer shares one precision, $b_l^{(w)} = b_l^{(a)} = b_l$, although the formulation allows the two to differ. We use the standard SAM 2 segmentation loss over the frames $t = 1, \ldots, T$ of a clip,

$$\mathcal{L}_{\text{seg}} = \sum_t \left[\lambda_f \mathcal{L}^t_{\text{focal}} + \lambda_d \mathcal{L}^t_{\text{dice}} + \lambda_{\text{iou}} \mathcal{L}^t_{\text{iou}} + \lambda_c \mathcal{L}^t_{\text{class}} \right], \tag{3}$$

and straight-through estimation gives

$$\frac{\partial \mathcal{L}_{\text{seg}}}{\partial W_l} \approx \frac{\partial \mathcal{L}_{\text{seg}}}{\partial W_l^{(q)}}, \qquad \frac{\partial \mathcal{L}_{\text{seg}}}{\partial b_l} \approx \frac{\partial \mathcal{L}_{\text{seg}}}{\partial Q} \cdot \frac{\partial Q}{\partial \Delta} \cdot \frac{\partial \Delta}{\partial b_l}. \tag{4}$$

Weights and precision are updated together. Where reducing precision at a layer barely moves the segmentation objective, the gradient pushes b_l down. Where a component is precision-critical, including output-facing heads, reducing b_l raises the loss and the profile keeps that layer higher.

We add two standard stabilizers, since minibatch gradients on b_l are noisy early in training and small changes in b_l move the quantization grid, and can produce oscillatory updates or collapse in a few layers. We clip each b_l to a valid interval, keeping the grid well defined and confining the profile to a usable range. We stabilize early precision updates with exponential moving-average smoothing and report the final precision parameters. Neither device is specific to this setting, and neither encodes a module-specific allocation rule.

3.3 SELECTION COST IN QUANTIZED SAM 2

Consider one *propagation unit*, an object at a frame where SAM 2 extends an existing mask without a new point prompt, so that $M > 1$ candidates are produced. Let $\{Z_m\}_{m=1}^M$ be the candidate mask logits, Y the ground-truth mask, and $\mathbf{s} \in \mathbb{R}^M$ the predicted-IoU scores. The *quality* q_m of candidate m is its soft IoU against Y with no threshold on the logits,

$$q_m = \frac{\sum \sigma(Z_m) \odot Y + \varepsilon}{\sum \sigma(Z_m) + \sum Y - \sum \sigma(Z_m) \odot Y + \varepsilon}, \tag{5}$$

Here σ is elementwise sigmoid, sums run over pixels, and larger $q_m \in [0, 1]$ means candidate m better agrees with Y. We detach $\mathbf{q}$ to avoid training masks toward this measure, and define oracle quality as $Q_{\text{oracle}} = \max_m q_m$.

At inference, $\hat{m} = \arg\max_m s_m$ obtains quality $q_{\hat{m}}$. A differentiable distribution $\mathbf{p}$, with $p_m = \text{softmax}(s_m/\tau)$ at temperature τ, passes gradients to the scores and yields expected quality $Q_{\text{sel}} = \sum_m p_m q_m$. Soft and hard *regret* measure quality forfeited relative to the oracle:

$$R = Q_{\text{oracle}} - Q_{\text{sel}}, \qquad R^{\text{hard}} = Q_{\text{oracle}} - q_{\hat{m}}. \tag{6}$$

The objective of the next subsection acts on R, while we report R^{hard}, the quality a selection error costs at inference. The *candidate-quality spread* is $\Delta q = \max_m q_m - \min_m q_m$. Because Δq shrinks as candidates collapse toward a common low quality, we also use $\widetilde{\Delta q} = \Delta q / \sqrt{\bar{q}\,(1-\bar{q})}$, with $\bar{q} = \frac{1}{M}\sum_m q_m$. An increase in R can reflect worse ranking ($\mathbf{p}$ moving off the oracle) or wider spread at fixed ranking. We distinguish *skill* using hard selection accuracy $\mathbf{1}[\hat{m} = \arg\max_m q_m]$, which has no spread term, and *stakes* using $\widetilde{\Delta q}$.

Let O, H, and $\bar{q}$ denote evaluation-set mean oracle, obtained, and candidate quality. Define *opportunity spread* $S = O - \bar{q}$, *selection gap* $G = O - H$, and *normalized skill* $K = (H - \bar{q})/S$. S compares an uninformed choice ($\bar{q}$) with the oracle (O), and G is the average R^{hard}. K is the share of S captured, but its normalization lets K change without a change in ordering. By construction, $G = S(1-K)$, with G the uncaptured opportunity. S uses oracle-minus-mean quality, whereas $\widetilde{\Delta q}$ compares the extremes.

Candidates differ in scale and fine-structure mass, and the sweep confirms uneven damage (Figure 1b). A fixed-slot policy always returns slot j and uses the head's relative cost: mean forfeited quality divided by mean oracle quality. An uninformed policy averages over the M slots, while the best fixed slot is chosen globally rather than per unit. From full precision to five bits, one slot's cost increases 5.1 times and the others 1.4 and 1.9 times. The worst full-precision slot is no longer worst at five bits, unlike uniform damage. The head forfeits less than every fixed slot at full precision but more than the best at five bits. Relative to the best slot, its cost is 0.603 and 2.091, with intervals $[0.403, 0.800]$ and $[1.433, 3.122]$. At six bits they are level, 1.017 with $[0.823, 1.243]$, so the head's advantage narrows monotonically with precision.

The account makes a prediction the learned profile tests. An allocation learned under the task loss, with no term scoring the comparison among candidates, should concentrate the allocated precision on the predicted-IoU head. Until candidate quality collapses, the rising selection cost runs through normalized skill, whose interaction with spread is not separated by these measurements. We therefore supervise the cost itself. Differentiating spread through $\mathbf{q}$ would train masks toward the soft IoU of Eq. (5), which detachment prevents. Because full precision also incurs regret, we penalize only the excess introduced by quantization.

3.4 Teacher-Relative Selection-Regret Objective

We add a term that penalizes that excess in regret and nothing else, above a margin γ, zero unless stated, averaged over the N eligible units in a batch. For a propagation unit i, let R_i^S and R_i^T be the soft regret of Eq. (6) for the quantized student and a frozen full-precision copy, both under the *same* conditioning mask and with R_i^T detached, giving

$$\mathcal{L}_{\text{SR}} = \frac{1}{N}\sum_{i=1}^{N} \left[\, R_i^S - R_i^T - \gamma \,\right]_+. \tag{7}$$

Eligibility is restricted to mask-prompted propagation frames with no point prompt and $M > 1$ present-object candidates, with teacher and student verified to share the conditioning mask. Other frames contribute nothing.

With $\mathbf{q}$, Q_{oracle}, and R^T detached, $\mathcal{L}_{\text{SR}}$ backpropagates only through the student's distribution $\mathbf{p}$, predicted-IoU head, and shared upstream decoder and quantization parameters. Detaching $\mathbf{q}$ blocks direct mask gradients within a step, but shared parameters allow candidate generation to change across training (Table 3). We combine the objectives with one fixed weight:

$$\mathcal{L} = \mathcal{L}_{\text{seg}} + \lambda_{\text{sr}}\,\mathcal{L}_{\text{SR}}. \tag{8}$$

The segmentation loss scores each candidate on its own through its per-candidate IoU regression, and $\mathcal{L}_{\text{SR}}$ adds the term defined on the comparison among the candidates.

4 Experiments

4.1 Experiment Settings

We follow the SAM 2 protocol (Ravi et al., 2025) and train on a joint image and video stream. The methods we train ourselves share initialization, training subset, prompts, and schedule, and differ

only in how precision is assigned and, for our full model, in the regret term. We adapt HAWQv2 and MixQSAM to our budget and run them on the same harness, but they assign precision from single-frame calibration.

4.1.1 DATASETS AND METRICS.

Full-scale SAM 2 training uses SA-1B (Kirillov et al., 2023) (11M images, 1B masks) and SA-V (Ravi et al., 2025) (51K videos, 643K masklets), impractical for repeated QAT. We therefore perform low-data QAT by sampling **1,000 images** and **100 videos** from the first training split of SA-1B and SA-V, fixed across all methods.

Video evaluation. We evaluate on the SA-V test split (150 videos). We use the first frame's ground-truth mask as the prompt and report $\mathcal{J}\&\mathcal{F}$ (Pont-Tuset et al., 2018).

Image evaluation. We evaluate on COCO2017 validation (Lin et al., 2014) (5,000 images, 36,335 instances). We report mIoU, using ground-truth bounding boxes as prompts.

Candidate-selection diagnostic. A unit is a propagation frame with a present object, more than one candidate, and no point prompt, so its candidates come from memory alone. One manifest of such units, from 49 of the 100 videos and shared by every checkpoint, carries every sweep row at a uniform width in place of the learned profile. The units are not held out, and every comparison is paired within them. We add the rank correlation τ_b between scores and quality over the units where it is defined at every reported precision, call candidates interchangeable at a smallest pairwise IoU of 0.95, and bootstrap by trajectory cluster.

4.2 IMPLEMENTATION

Experimental Setup. We quantize the Tiny, Small, and Base+ variants of SAM 2.1 from their full-precision pretrained weights, leaving the topology unchanged and inserting fake-quantization modules at the convolution and linear operators of the image encoder, the memory attention, the memory encoder, the prompt encoder, and the mask decoder. All layers start at 6 bits and are clipped to $[2, 8]$, and we report the parameter-weighted average of the final precision values as the operating point, continuous since b_l is.

Baselines. We train MinMax QAT (Jacob et al., 2018) at fixed W6A6, W5A5, and W4A4, run PTQ at W4, and apply HAWQv2 (Dong et al., 2020) and MixQSAM (Ranjan & Savakis, 2025) to SAM 2. Numbers for Q-SAM2 (Farronato et al., 2025) and MixQSAM2 (Duan et al., 2026) are quoted from their papers and not protocol-matched.

Objective settings. The regret term of Eq. (8) is weighted at $\lambda_{\mathrm{sr}} = 1.64$, the soft regret uses temperature $\tau = 0.2$, and the margin of Eq. (7) is $\gamma = 0$ throughout. The training configuration is in the Appendix.

System. We use PyTorch 2.6.0 with CUDA 11.8 on a single NVIDIA A6000 48GB GPU, disable AMP, and train in FP32 so that gradients through the bit parameters stay stable.

4.3 SEGMENTATION PERFORMANCE

Table 1 compares the learned profile with full precision, uniform PTQ and QAT, and mixed-precision baselines at three model scales. At under five bits on average, it loses 6.6, 3.3, and 1.9 SA-V points against full precision on Base+, Small, and Tiny, respectively, while staying within 0.014 COCO mIoU. It avoids uniform W4A4 collapse on Base+ and Tiny. Base+ exceeds W6A6 by 11.1 points at a lower budget, while uniform quantization already reaches 73.3 and 74.4 on Small and Tiny. At the same average budget, both mixed-precision baselines stay within 0.002 COCO mIoU of full precision on Base+ but fall to low-single-digit SA-V scores on Base+ and Tiny, while ours remains above 71 at every scale. In this comparison, allocations based on single-frame calibration did not preserve video performance. Quoted Q-SAM2 uses a tighter budget and its own protocol, with lower scores at every scale. MixQSAM2 reports only one scale on a different backbone.

Table 1: Segmentation performance comparison. Bold marks our rows.

Models	Method	Bitwidth	SA-V test ($\mathcal{J}\&\mathcal{F}\uparrow$) ($\mathcal{J}, \mathcal{F}$)	COCO2017 val (mIoU ↑)
SAM 2.1 Base+	FP	FP32	77.67 (74.24, 81.10)	0.7568
	PTQ (Jacob et al., 2018)	W4	48.72 (46.46, 50.99)	0.5757
		W6A6	59.97 (56.13, 63.80)	0.7305
	MinMax (Jacob et al., 2018)	W5A5	55.06 (51.10, 59.01)	0.7262
		W4A4	41.97 (38.21, 45.72)	0.5787
	HAWQv2 (Dong et al., 2020)	W4.80 (Avg.)	2.14 (2.18, 2.09)	0.7567
	MixQSAM (Ranjan & Savakis, 2025)	W4.80 (Avg.)	2.32 (1.80, 2.85)	0.7549
	Q-SAM2† (Farronato et al., 2025)	W2A4	64.30 (-, -)	0.4940
	MixQSAM2† (Duan et al., 2026)	Avg-4	50.30 (-, -)	-
	Ours	**W/A 4.79 (Avg.)**	**71.09 (67.35, 74.82)**	**0.7433**
SAM 2.1 Small	FP	FP32	76.92 (73.31, 80.52)	0.7581
	PTQ	W4	26.90 (27.01, 26.79)	0.6020
		W6A6	74.97 (71.16, 78.79)	0.7529
	MinMax	W5A5	73.68 (69.60, 77.77)	0.7418
		W4A4	73.25 (69.41, 77.09)	0.7525
	HAWQv2	W4.80 (Avg.)	25.33 (25.33, 25.33)	0.7139
	MixQSAM	W4.80 (Avg.)	11.62 (11.49, 11.75)	0.7217
	Q-SAM2†	W2A4	67.90 (-, -)	0.4730
	Ours	**W/A 4.79 (Avg.)**	**73.62 (69.99, 77.26)**	**0.7469**
SAM 2.1 Tiny	FP	FP32	76.54 (73.02, 80.07)	0.7662
	PTQ	W4	31.33 (28.71, 33.95)	0.6514
		W6A6	75.23 (71.36, 79.12)	0.7600
	MinMax	W5A5	74.43 (70.56, 78.29)	0.7636
		W4A4	32.25 (29.68, 34.83)	0.6112
	HAWQv2	W4.80 (Avg.)	1.74 (1.66, 1.82)	0.7257
	MixQSAM	W4.80 (Avg.)	2.79 (1.54, 4.04)	0.4427
	Q-SAM2†	W2A4	67.10 (-, -)	0.4750
	Ours	**W/A 4.80 (Avg.)**	**74.68 (70.96, 78.40)**	**0.7557**

† Quoted from the original paper. MixQSAM2 reports SAM 2 base only, so no SAM 2.1 Small or Tiny entry exists.

Table 2: Precision sweep on one Base+ checkpoint using the same SA-V propagation units. Orig. denotes pretrained SAM 2.1. Selection accuracy degrades before oracle candidate quality; metrics follow Section 3.3.

Prec.	$O\uparrow$	$G\downarrow$	S	K	Acc↑	τ_b	Intchr.
Orig.	0.8951	0.0098	0.0302	0.674	0.597	$+0.537$	-
FP32	0.8900	0.0073	0.0356	0.794	0.654	$+0.577$	47.2%
8-bit	0.8864	0.0114	0.0350	0.675	0.563	$+0.519$	49.4%
6-bit	0.8695	0.0165	0.0353	0.532	0.485	$+0.518$	39.4%
5-bit	0.7067	0.0376	0.0684	0.450	0.403	$+0.476$	6.5%
4-bit	0.0593	0.0065	0.0096	0.319	0.385	-	12.1%

Table 3: Selection-regret ablation on Base+ at matched average precision. Each variant is trained separately and evaluated on the same propagation units. $\bar{q}$, O, and H denote mean candidate, oracle, and selected-mask quality; bold marks their highest values.

Config	Bits	$\bar{q}\uparrow$	$O\uparrow$	$H\uparrow$	Acc↑	$\mathcal{J}\&\mathcal{F}$	mIoU
Prec.-only	4.791	0.5697	0.6534	0.6260	0.558	70.47	0.7344
Top-tail	4.792	0.5754	0.6606	0.6323	0.558	71.34	0.7356
Absolute	4.792	0.6014	0.6744	0.6555	0.571	70.04	0.7365
Ours	4.791	**0.6230**	**0.6943**	**0.6723**	0.563	71.09	0.7433

4.4 Qualitative Evaluation

MixQSAM masks fragment during video propagation, while the learned profile tracks the target (Figure 4). MixQSAM also leaks and roughens boundaries under static single-point prompts (Figure 5), exposing ambiguity removed by boxes. The uniform MinMax baseline achieves comparable stability only at the higher W6A6 budget.

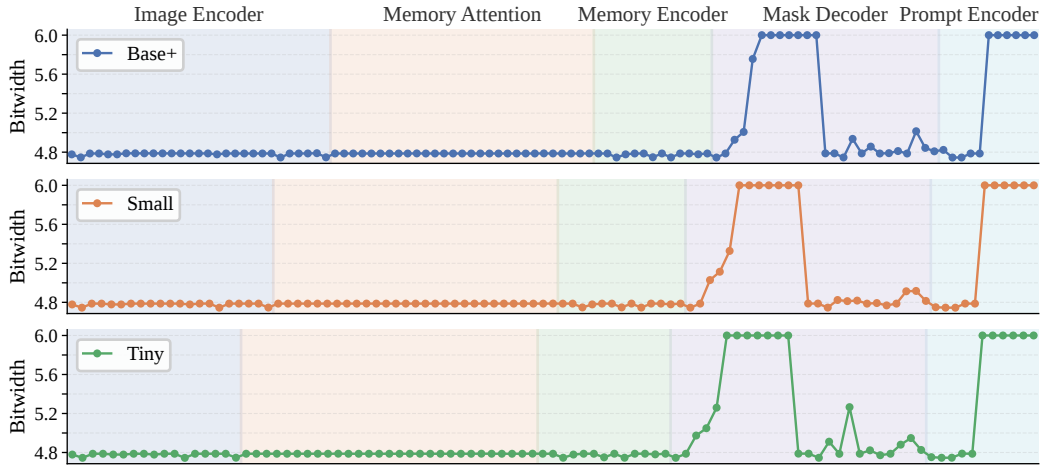

Figure 3: Final per-layer bit allocation map at three model scales. Most layers settle into a narrow low-bit region, and among those the controller lowers the highest lie in the mask decoder, the single highest on the predicted-IoU readout.

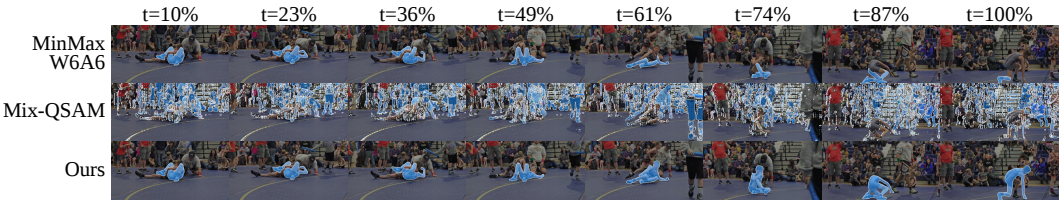

Figure 4: Visualization of video segmentation under quantization (SA-V, Base+). One clip, prompted by the first-frame mask.

4.5 Selection Cost across Precisions

We execute one trained Base+ checkpoint at uniform 8, 6, 5, and 4 bits (Table 2). The $M = 3$ candidates give an uninformed selector one-third accuracy. Eight bits exceeds every learned layer precision and six bits is never lower, so the loss is not caused by stripping away the learned allocation. At full precision, fine-tuning improved every selection statistic over the original SAM 2.1. At six bits, oracle quality loses 2.3% but the hit rate drops a quarter (-16.9 points, paired bootstrap interval $[-24.8, -9.0]$), while relative selection cost rises 2.3 times ($[1.43, 3.99]$). At five bits, selection cost rises 6.4 times ($[3.97, 11.96]$), oracle quality loses 21%, and the interchangeable share falls from 47.2% to 6.5%.

Hard selection accuracy falls significantly at eight, six, and five bits, while the rank correlation declines less, so the selector loses the top of its ordering faster than the ordering itself. Decomposing the rise in the selection gap through $G = S(1 - K)$ splits the four working precisions in two. Through six bits the opportunity spread does not move, and the whole rise comes from the normalized skill K, $+0.0093$ of an observed $+0.0092$. At five bits the candidate set reorganizes, and the rise divides into $+0.0068$ from the widening spread, $+0.0122$ from K, and $+0.0113$ from their interaction.

At five bits, within-trajectory normalized spread $\widetilde{\Delta q}$ grows with propagation depth (Figure 1c): $+0.1330$ from the first propagated frame to the third across 58 trajectories, with interval $[+0.0648, +0.2049]$. Neither the full-precision control on the same trajectories nor still-image training clips show this rise. Selection accounts for 53%, 31%, and 14% of the hard-selected quality drop at eight, six, and five bits, respectively, dominating first before generation takes over. At four bits, oracle quality collapses to 0.059, leaving little quality to forfeit and shrinking the gap, so we do not decompose this row. The interchangeable share rises only through pixel-identical zero-quality candidates with no best candidate.

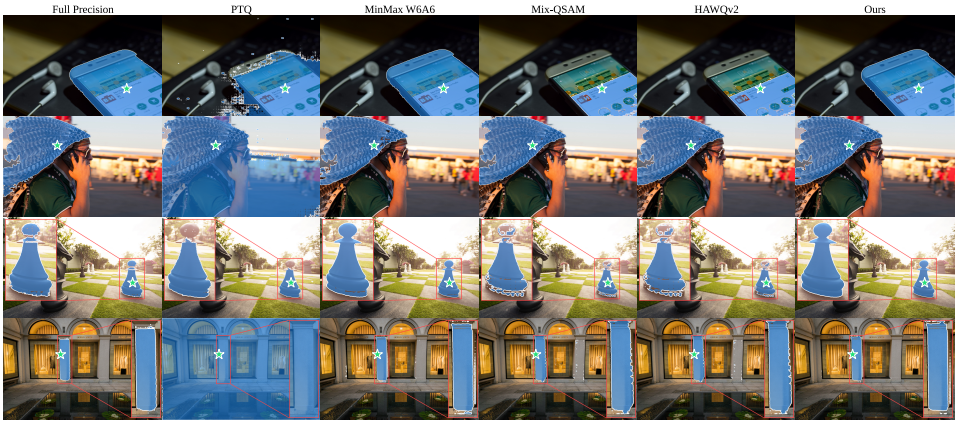

Figure 5: Single-point image segmentation (SA-1B, Base+). Masks for one positive point.

4.6 Where Learned Precision Concentrates

The profile learned under $\mathcal{L}_{\text{seg}}$ alone in Figure 3 is separate from the full-objective **Ours** rows of Table 1. Among the layers lowered by the controller, the last predicted-IoU layer ranks highest, and the preceding two rank second and fourth. The segmentation loss supervises both this readout and mask decoding, yet upscaling and hypernetwork layers remain below 4.79 bits while the final readout layer holds 5.85. The sweep and learned allocation identify the same three layers through separate measurements.

4.7 Effect of the Selection-Regret Objective

We add the selection-regret term at a matched 4.79-bit budget and report it in Table 3. Every row trains its own profile and is measured on the units of Table 2, so the two tables share units but not precision. The term raises the candidate mean by $+0.0533$, the oracle by $+0.0409$, and the obtained quality by $+0.0463$, each with a paired interval excluding zero. End-task scores are close to the precision-only baseline on SA-V and COCO.

Two variants isolate the objective's components. Averaging only the worst $\lceil \rho N \rceil$ units at tail fraction ρ leaves the single worst unit at the available N and produces under a fifth of the default's movement on all three candidate statistics. Thus aggregation controls whether the candidate set changes substantially. Penalizing absolute rather than teacher-relative regret retains most of the movement but finishes lower by 0.0216, 0.0199, and 0.0168, with intervals excluding zero. The teacher baseline therefore affects the quality obtained. Ours leads on all three candidate statistics and COCO, while the four rows span only 1.3 points of $\mathcal{J}\&\mathcal{F}$. The Appendix reports candidate-level sensitivity to regret weight, temperature, tail fraction, and eligibility.

5 Conclusion

Selection degrades before the masks do in quantized SAM 2. Hard selection accuracy falls significantly at eight bits while candidate quality remains within half a percent of full precision. Candidate quality gives way at five bits, with uneven slot damage and increasing effects across propagation depth. The segmentation-loss-only profile peaks at the same predicted-IoU readout. Teacher-relative selection regret raises candidate-set quality at a matched budget without moving the spread, gap, or accuracy. At 4.79 bits, the learned profile retains 71.1 SA-V $\mathcal{J}\&\mathcal{F}$ while two matched-budget mixed-precision baselines fall below 2.4. Why top-choice accuracy degrades faster than overall ranking, and whether this account extends to other models that score their own candidates, remain open as future work.

AI USE STATEMENT

We designed the experiments ourselves and did not use AI to implement the core method. For coding tasks, AI assistance was limited to experiment automation: running sequences of experiments and launching pre-planned follow-up experiments based on the preceding results. We also used generative AI tools for translation, drafting and revising portions of the manuscript, language polishing, LaTeX formatting, and figure placement. AI tools provided advice on interpreting results and examining reviewer feedback; we made the final decisions about result interpretation and the manuscript's scientific content. We reviewed and double-checked AI-assisted code and text, verified and corrected the automation code, and checked text revisions for technical accuracy and preservation of the intended meaning. We take responsibility for the final content of this work, including AI-assisted text and code.

ETHICS STATEMENT

We study quantization for image and video segmentation using the existing SA-1B, SA-V, and COCO2017 datasets. These datasets include images of people and everyday environments, so use of the data and resulting models should respect privacy and applicable data and model licenses.

REPRODUCIBILITY STATEMENT

Section 3 specifies the quantization parameterization, precision-learning procedure, and selection-regret objective. Sections 4.1 and 4.2 describe the experimental and implementation settings. The appendix details the datasets, subsets, prompts, and metrics (Section A.1), hyperparameters and compute environment (Section A.2), quantization scope and controller settings (Sections A.3 and A.4), and seed variability (Section D.7). Table 16 documents the baseline adaptations and their calibration limitations.

REFERENCES

Ron Banner, Yury Nahshan, and Daniel Soudry. Post training 4-bit quantization of convolutional networks for rapid-deployment. *Advances in neural information processing systems*, 32, 2019.

Yash Bhalgat, Jinwon Lee, Markus Nagel, Tijmen Blankevoort, and Nojun Kwak. Lsq+: Improving low-bit quantization through learnable offsets and better initialization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 696–697, 2020.

Yu Cheng, Duo Wang, Pan Zhou, and Tao Zhang. A survey of model compression and acceleration for deep neural networks. *arXiv preprint arXiv:1710.09282*, 2017.

Jungwook Choi, Zhuo Wang, Swagath Venkataramani, Pierce I-Jen Chuang, Vijayalakshmi Srinivasan, and Kailash Gopalakrishnan. Pact: Parameterized clipping activation for quantized neural networks. *arXiv preprint arXiv:1805.06085*, 2018.

Yoni Choukroun, Eli Kravchik, Fan Yang, and Pavel Kisilev. Low-bit quantization of neural networks for efficient inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 3009–3018, 2019. doi: 10.1109/ICCVW.2019.00363.

Zhen Dong, Zhewei Yao, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Hawq: Hessian aware quantization of neural networks with mixed-precision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 293–302, 2019.

Zhen Dong, Zhewei Yao, Daiyaan Arfeen, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Hawq-v2: Hessian aware trace-weighted quantization of neural networks. *Advances in neural information processing systems*, 33:18518–18529, 2020.

Yuzhe Duan, Xuanxuan Ren, Guizhe Dong, Xu Yang, and Yanhua Yang. Mix-qsam2: Mixed-precision quantization for high fidelity segmentation in resource constrained scenarios. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 3741–3749, 2026.

Nicola Farronato, Florian Scheidegger, Mattia Rigotti, Cristiano Malossi, Michele Magno, and Haotong Qin. Q-sam2: Accurate quantization for segment anything model 2. *arXiv preprint arXiv:2506.09782*, 2025.

Eran Goldman, Roei Herzig, Aviv Eisenschtat, Jacob Goldberger, and Tal Hassner. Precise detection in densely packed scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5227–5236, 2019.

Ruihao Gong, Xianglong Liu, Shenghu Jiang, Tianxiang Li, Peng Hu, Jiazhen Lin, Fengwei Yu, and Junjie Yan. Differentiable soft quantization: Bridging full-precision and low-bit neural networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4852–4861, 2019.

Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 3779–3787, 2019.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2704–2713, 2018.

Jangho Kim, Yash Bhalgat, Jinwon Lee, Chirag Patel, and Nojun Kwak. Qkd: Quantization-aware knowledge distillation. *arXiv.org*, 2019.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Darryl Lin, Sachin Talathi, and Sreekanth Annapureddy. Fixed point quantization of deep convolutional networks. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2849–2858. PMLR, 2016.

Tsung-Yi Lin, M. Maire, Serge J. Belongie, James Hays, P. Perona, Deva Ramanan, Piotr Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014.

Xiaoyu Liu, Xin Ding, Lei Yu, Yuanyuan Xi, et al. Pq-sam: Post-training quantization for segment anything model. In *European Conference on Computer Vision (ECCV)*, 2024.

Christos Louizos, Matthias Reisser, Tijmen Blankevoort, Efstratios Gavves, and Max Welling. Relaxed quantization for discretized neural networks. In *International Conference on Learning Representations*, 2019.

Haisheng Lu, Yujie Fu, Fan Zhang, and Le Zhang. Efficient quantization-aware training on segment anything model in medical images and its deployment. In Jun Ma, Yuyin Zhou, and Bo Wang (eds.), *Medical Image Segmentation Foundation Models. CVPR 2024 Challenge: Segment Anything in Medical Images on Laptop*, pp. 137–150, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-81854-7.

Chengtao Lv, Hong Chen, Jinyang Guo, Yifu Ding, and Xianglong Liu. Ptq4sam: Post-training quantization for segment anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Asit Mishra and Debbie Marr. Apprentice: Using knowledge distillation techniques to improve low-precision network accuracy. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=B1ae1lZRb`.

Markus Nagel, Marios Fournarakis, Rana Ali Amjad, Yelysei Bondarenko, Mart Van Baalen, and Tijmen Blankevoort. A white paper on neural network quantization. *arXiv preprint arXiv:2106.08295*, 2021.

Miloš Nikolić, Ghouthi Boukli Hacene, Ciaran Bannon, Alberto Delmas Lascorz, Matthieu Courbariaux, Omar Mohamed Awad, Isak Edo Vivancos, Yoshua Bengio, Vincent Gripon, and Andreas Moshovos. Bitpruning: Learning bitlengths for aggressive and accurate quantization. In *2024 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–5. IEEE, 2024.

Antonio Polino, Razvan Pascanu, and Dan Alistarh. Model compression via distillation and quantization. *arXiv preprint arXiv:1802.05668*, 2018.

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation, 2018. URL https://arxiv.org/abs/1704.00675.

Navin Ranjan and Andreas Savakis. Mix-qsam: Mixed-precision quantization of the segment anything model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 3305–3315, June 2025.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollar, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/45c1f6a8cbf2da59ebf2c802b4f742cd-Paper-Conference.pdf.

Xuanxuan Ren, Xiangyu Li, Kun Wei, Xu Yang, and Yanhua Yang. Q-minisam2: A quantization-based benchmark for resource-efficient video segmentation. In James Kwok (ed.), *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 1829–1837. International Joint Conferences on Artificial Intelligence Organization, 8 2025. doi: 10.24963/ijcai.2025/204. URL https://doi.org/10.24963/ijcai.2025/204. Main Track.

Pierre Stock, Armand Joulin, Rémi Gribonval, Benjamin Graham, and Hervé Jégou. And the bit goes down: Revisiting the quantization of neural networks. *arXiv preprint arXiv:1907.05686*, 2019.

Chen Tang, Kai Ouyang, Zhi Wang, Yifei Zhu, Wen Ji, Yaowei Wang, and Wenwu Zhu. Mixed-precision neural network quantization via learned layer-wise importance. In *European conference on computer vision*, pp. 259–275. Springer, 2022.

Stefan Uhlich, Lukas Mauch, Fabien Cardinaux, Kazuki Yoshiyama, Javier Alonso Garcia, Stephen Tiedemann, Thomas Kemp, and Akira Nakamura. Mixed precision dnns: All you need is a good parametrization. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Hyx0slrFvH.

Kuan Wang, Zhijian Liu, Yujun Lin, Ji Lin, and Song Han. Haq: Hardware-aware automated quantization with mixed precision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8612–8620, 2019.

Zixun Wang and Ben Dai. Rankseg-rma: An efficient segmentation algorithm via reciprocal moment approximation. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 163310–163333. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/ef1cace6eccc84ea1993db63edbad7d6-Paper-Conference.pdf.

Bichen Wu, Yanghan Wang, Peizhao Zhang, Yuandong Tian, Peter Vajda, and Kurt Keutzer. Mixed precision quantization of convnets via differentiable neural architecture search. *arXiv preprint arXiv:1812.00090*, 2018.

Linjie Yang and Qing Jin. Fracbits: Mixed precision quantization via fractional bit-widths. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10612–10620, 2021.

Zhewei Yao, Zhen Dong, Zhangcheng Zheng, Amir Gholami, Jiali Yu, Eric Tan, Leyuan Wang, Qijing Huang, Yida Wang, Michael Mahoney, et al. Hawq-v3: Dyadic neural network quantization. In *International Conference on Machine Learning*, pp. 11875–11886. PMLR, 2021.

Jing Zhang, Zhikai Li, Chengzhi Hu, Xuewen Liu, and Qingyi Gu. Saq-sam: Semantically-aligned quantization for segment anything model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 12547–12555, 2026.

Wenlun Zhang, Yunshan Zhong, Shimpei Ando, and Kentaro Yoshioka. Ahcptq: Accurate and hardware-compatible post-training quantization for segment anything model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22383–22392, October 2025.