SAQ-SAM: Semantically-Aligned Quantization for Segment Anything Model

Jing Zhang[1, 2], **Zhikai Li**[1,*], **Chengzhi Hu**[1, 2], **Xuewen Liu**[1, 2], **Qingyi Gu**[1,*]
[1]Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
{zhangjing2024, lizhikai2020, huchengzhi2024, liuxuewen2023, qingyi.gu}@ia.ac.cn

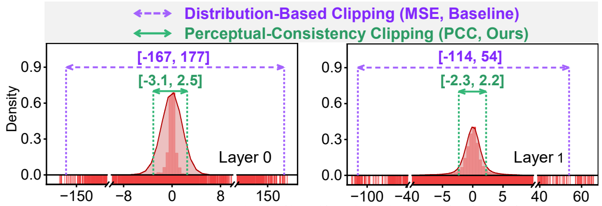
Abstract

Segment Anything Model (SAM) exhibits remarkable zero-shot segmentation capability; however, its prohibitive computational costs make edge deployment challenging. Although post-training quantization (PTQ) offers a promising compression solution, existing methods yield unsatisfactory results when applied to SAM, owing to its specialized model components and promptable workflow: (i) The mask decoder's attention exhibits extreme activation outliers, and we find that aggressive clipping (even 100x), without smoothing or isolation, is effective in suppressing outliers while maintaining performance. Unfortunately, traditional distribution-based metrics (e.g., MSE) fail to provide such large-scale clipping. (ii) Existing quantization reconstruction methods neglect semantic interactivity of SAM, leading to misalignment between image feature and prompt intention. To address the above issues, we propose SAQ-SAM in this paper, which boosts PTQ for SAM from the perspective of semantic alignment. Specifically, we propose Perceptual-Consistency Clipping, which exploits attention focus overlap to promote aggressive clipping while preserving semantic capabilities. Furthermore, we propose Prompt-Aware Reconstruction, which incorporates image-prompt interactions by leveraging cross-attention in mask decoder, thus facilitating alignment in both distribution and semantic. Moreover, to ensure the interaction efficiency, we design a layer-skipping strategy for image tokens in encoder. Extensive experiments are conducted on various SAM sizes and tasks, including instance segmentation, oriented object detection, and semantic segmentation, and the results show that our method consistently exhibits advantages. For example, when quantizing SAM-B to 4-bit, SAQ-SAM achieves 11.7% higher mAP than the baseline in instance segmentation task.

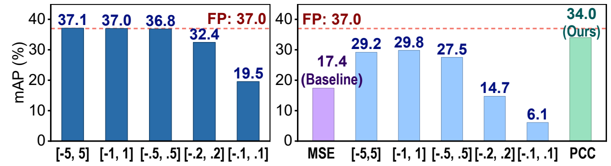
Code — https://github.com/jingjing0419/SAQ-SAM

1 Introduction

Segment Anything Model (Kirillov et al. 2023) (SAM) shows promising applications as the base model for promptable segmentation. Benefiting from sufficient pre-training and prompt-guided fine-tuning, SAM exhibits strong zero- or few-shot generalization capability to interactively segment regions of interest to the user. However, SAM's powerful representation capability is accompanied by a large number of parameters and high computational costs, limiting its potential on resource-constrained devices (Zhang, Cai, and Han 2024; Chen et al. 2023).



(a) Extreme activation distribution in mask decoder



(b) FP with large-scale clipping (c) W6A6 with different clipping

Figure 1: Visualization of extreme activation distributions in mask decoder and the performance of different clipping methods. QK activations in the mask decoder show highly skewed distributions, with most data concentrated in a narrow range while outliers can exceed 180 times the normal range. MSE provides an overly wide clipping range, whereas our Perceptual-Consistency Clipping (PCC) method can identify outliers more precisely.

Model quantization (Li et al. 2024b; Liu et al. 2024b; Xiao et al. 2024) can reduce computational overhead and model size by converting weights and activations to low-bit integers (Li and Gu 2023; Gholami et al. 2021; Choi et al. 2018), and post-training quantization (PTQ) (Nagel et al. 2020; Li et al. 2023a), which efficiently calibrates quantization parameters using a small set of unlabeled data, stands out as a promising approach. Unfortunately, SAM features the unique activation distribution and network architecture, which pose new challenges to PTQ, rendering conventional methods inadequate. To this end, several studies (Li, Zhang, and Gu 2024) have attempted to propose specific solutions

---

[*] Corresponding authors.

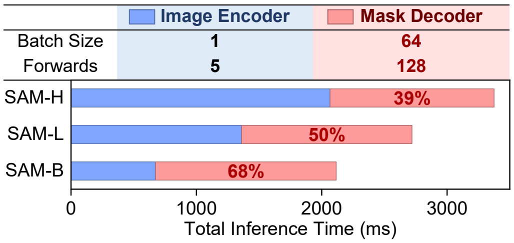


Figure 2: Comparison of total inference time between image encoder and mask decoder in semantic segmentation task.

for SAM's characteristics. For instance, PQ-SAM (Liu et al. 2024a) hierarchically clusters similar channels to learn unified transformation factors. However, it is limited to the image encoder with a classical Transformer structure and lacks adaptability to the mask decoder, which features more complex connection paths. In practice, although SAM's parameters are primarily concentrated in the image encoder, the mask decoder can also introduce significant computational overhead. As shown in Figure 2, in semantic segmentation task, given a single image, the mask decoder must perform multiple large-batch inferences for numerous prompting points, whereas the image encoder requires much fewer forward passes (Chen, Yang, and Zhang 2023). To this end, PTQ4SAM (Lv et al. 2024) is proposed for quantization bottleneck in the mask decoder, which merges bimodal distributions in post-key-linear activations through QK weight transformation. However, it still seeks to align full-precision (FP) model at distribution level, leading to unsatisfactory performance, especially for low-bit quantization.

In this paper, we comprehensively investigate SAM's model architecture and interactivity in promptable segmentation, and find that the semantic misalignment constitutes the main challenge for low-bit quantization. In the following, we will discuss the misalignment issues in two critical techniques of PTQ: outlier handling and reconstruction.

i) Semantic Misalignment in Outlier Handling. QK activations in SAM's mask decoder exhibits extreme outliers, as shown in Figure 1(a). Fortunately, we find an interesting phenomenon: aggressively clipping these outliers has little impact on final segmentation performance. As shown in Figure 1(b), clipping the skewed distributed activation from [-167, 177] and [-114, 54] to [-1, 1] has little impact. More importantly, such large-scale range reduction improves quantization resolution, thereby reducing quantization errors. To this end, we perform a grid search of large-scale clipping intervals for quantization calibration, as shown in Figure 1(c), where [-1, 1] achieves 12.4% performance improvement compared with traditional MSE metric. Thus, distribution-based metrics are no longer applicable to segmentation tasks, resulting in severe semantic degradation, as shown in Figure 3. As a result, it is crucial to establish a semantic level metric to drive optimal-scale clipping.

ii) Semantic Misalignment in Reconstruction. Quantization reconstruction is a prevalent method for restoring model performance. However, existing methods typically target at simple tasks with single branch model architecture, thus exhibiting suboptimal performance in promptable segmentation task with encoder-decoder architecture. This is because, locally reconstructing the FP responses in the image encoder overlooks the semantic intent of prompts, thus causing feature distortions in visual-prompt interactions and ultimately compromising segmentation accuracy. To this end, to ensure quantization accuracy, it is crucial to facilitate semantic alignment during reconstruction.

Figure 3: Attention heatmaps in the mask decoder with different quantization clipping methods. The distribution-aligned MSE leads to significant attention degradation, whereas our semantic-aligned PCC maintains the consistency with the FP model.

With the above insights, we propose SAQ-SAM, which significantly improves PTQ performance for SAM by promoting semantic-level alignment with the FP model. Specifically, for outlier suppression, we propose Perceptual-Consistency Clipping (PCC) for QK activations. By proactively measuring the quantization-induced deviation in attention focus, PCC achieves magnitude-independent outlier suppression with semantic alignment. As illustrated in Figure 3, PCC evidently mitigates the attention functional degradation issue. Further, we propose Prompt-Aware Reconstruction (PAR), which learns quantization parameters under the guidance of prompts. By leveraging SAM's off-the-shell mask decoder to facilitate interaction between prompts and image tokens, PAR reconstructs the interaction responses supervised by the FP model. This preserves the correspondence between visual features and prompts, thus facilitating alignment in both distribution and semantic. The main contributions are summarized as follows:

- We observe that aggressively clipping outliers in SAM has minimal impact on segmentation performance. With this insight, we propose Perceptual-Consistency Clipping, which utilizes attention focus deviation as a metric to guide large-scale clipping, and thus suppressing outliers while maintaining semantic alignment.
- We propose Prompt-Aware Reconstruction, which integrates image-prompt interactions into reconstruction. This enables the quantization model to align with the FP model in both feature modeling and prompt following. We also design a layer-skipping strategy of image tokens to ensure interaction efficiency.

- We conduct extensive experiments on SAMs of various sizes across three mainstream tasks, demonstrating the consistent superiority over baseline methods. For instance, in instance segmentation, our 4-bit SAQ-SAM achieves an average 14% mAP improvement on SAM-B and maintains nearly lossless accuracy on SAM-L.

2 Related Work

Segment Anything. Segment Anything (SAM) (Kirillov et al. 2023; Ravi et al. 2024) has emerged as a versatile and promptable image segmentation tool. Pre-trained on an extensive dataset, SAM demonstrates remarkable generalization across various downstream tasks, including instance segmentation (Chen et al. 2024), semantic segmentation (Chen, Yang, and Zhang 2023), and oriented object detection (Yu et al. 2023). Its capabilities extend to critical domains such as medical image annotation (Ma et al. 2024; Cheng et al. 2023), where it shows significant promise as a powerful diagnostic support tool. Thanks to its excellent generalization capabilities, SAM is also be used as an auxiliary branch for industrial anomaly detection (Li, Qi, and Geng 2025). However, SAM's powerful capabilities come with extensive memory and computational costs. Although some studies have proposed various lightweight variants to accelerate inference (Zhao et al. 2023; Zhang et al. 2023; Xiong et al. 2024), the model parameters of these methods are still kept in floating-point type. This prevents them from benefiting from efficient low-precision integer operation units, resulting in suboptimal efficiency.

Post-Training Quantization. Model quantization (Li et al. 2023b, 2022b; Liu, Li, and Gu 2025) reduces memory and computation costs by converting floating-point values to integers. To mitigate the performance degradation caused by precision loss, researchers have made various efforts, among which PTQ efficiently uses a handful of unlabeled samples to set quantization parameters (Lin et al. 2024; Liu, Li, and Gu 2024; Li et al. 2022c).

The backbone of SAM is built on the Transformer architecture, a structure for which PTQ have been extensively researched and developed. SmoothQuant (Xiao et al. 2023) employs equivalent transformations to smooth the activation distribution, mitigating the impact of harmful outliers. OmniQuant (Shao et al. 2023) improves the applicability of smoothing methods by learning equivalent transformation factors and clipping ranges. RepQ-ViT (Li et al. 2023c) and RepQuant (Li et al. 2024a) decouple quantization from inference, bridging the gap through scale reparameterization for efficient hardware-friendly quantization. In addition to these distribution-adjustment methods, some methods exploit the characteristics of the attention mechanism. For example, PTQ-ViT (Liu et al. 2021) determines quantization clipping by maintaining the consistency of the attention rank. Other methods, such as BRECQ (Li et al. 2021), QDrop (Wei et al. 2022), and PD-Quant (Liu et al. 2023), aim to learn optimal quantization parameters. They reconstruct the internal response of the quantization model under the supervision of the FP model.

Despite the aforementioned methods achieved remarkable performance in dealing with classical Transformer structures, SAM's distinctive activation distribution poses new challenges, leaving their performance subpar. To this end, several works try to make improvements. To tackle the channel-wise distribution imbalance in activations, PQ-SAM (Liu et al. 2024a) hierarchically clusters channels with similar distributions and learns shared transformation factors for each group, thereby reducing optimization complexity. PTQ4SAM (Lv et al. 2024) identify and integrate the bimodal distribution in post-key-linear activation in mask decoder, and implement an adaptive granularity quantizer for post-softmax activation to adapt to different types of attention module. However, these distribution-level improvements have shown unsatisfactory performance in practice.

3 Method

3.1 Preliminaries

Transformer Layers in SAM: In SAM's image encoder, the Transformer layers employ window attention and global attention alternately (Li et al. 2022a), and the former divides the image into non-overlapping windows and computes self-attention within each window without shifting. In the mask decoder, a lightweight Two-Way Transformer is utilized to update both the image embedding and prompt tokens via cross-attention, facilitating the information interaction between prompt and image. Specifically, the token-to-image cross-attention utilizes prompt tokens as queries and employs image tokens as keys and values, whereas the image-to-token cross-attention implements the inverse configuration to ensure comprehensive feature interaction.

Quantization Calibration: The uniform quantizer is the most commonly used and deployment-friendly quantizer, which is defined as:

$$\text{Quant}: x_q = \text{clip}\left(\left\lfloor \frac{x}{s} \right\rceil + z, 0, 2^b - 1\right), \tag{1}$$

$$\text{DeQuant}: \hat{x} = s\left(x_q - z\right) \approx x, \tag{2}$$

where x and x_q are the floating-point and quantized values, respectively. The dequantized value $\hat{x}$ approximates x. $\lfloor \cdot \rceil$ is the round-to-nearest operation. clip function truncates values outside the b-bit range. The quantization scale s and zero-point z are PTQ parameters to be searched, determined by the clipping boundaries x_{low} and x_{up} as follows:

$$s = \frac{x_{up} - x_{low}}{2^b - 1}, z = \left\lfloor -\frac{x_{low}}{s} \right\rceil. \tag{3}$$

The process of determining clipping boundaries is crucial for PTQ performance, which is also called calibration.

3.2 Perceptual-Consistency Clipping

Insight. With the observation that aggressively clipping the extreme outlier in QK activations, although significantly modifying the distribution, doesn't hurt the segmentation performance, we aim to break through the limitations of distribution alignment and turn to leverage the semantic nature of attention mechanisms for PTQ calibration. As stated in (Vaswani et al. 2017), in Transformers architecture, the attention mechanism captures the semantic information perceived by the model. It allocates greater focus to regions of

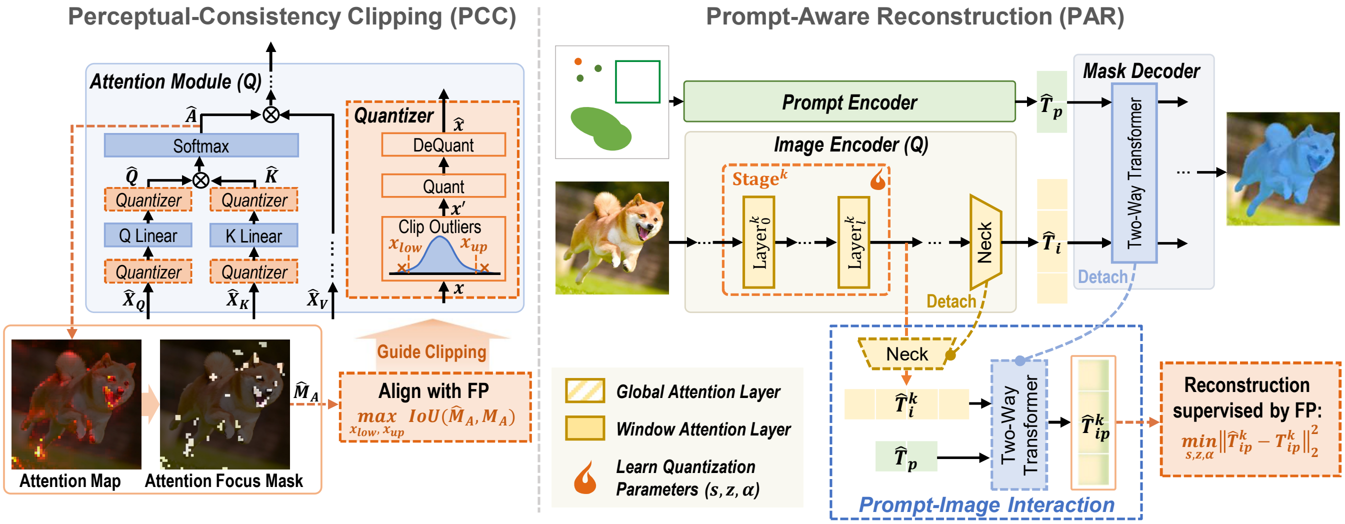


Figure 4: Overview of SAQ-SAM. The proposed PCC guides quantization clipping of QK activation by minimizing the Attention Focus deviation from FP, thereby semantically preserving the perceptual alignment. Our PAR incorporates image-prompt interactions into per-stage reconstruction, utilizing the off-the-shell module in the mask decoder. Through minimizing the interaction response error supervised by the FP model, quantization model learns correspondence between visual features and prompt intentions, thus facilitating dual alignment at both the distributional and semantic levels.

interest, thereby reinforcing the features most relevant to the task. Building on this, it's reasonable to leverage the potential relationships modeled by attention to preserve semantic-level consistency before and after quantization.

Attention Focus Overlap Metric. Inspired by above insights, we define highly attented region as the perceptual focus and maximize the overlap of the perceptual focus before and after quantization, thereby maintaining consistency in attention perception. The procedure of the metric formulation is described as follows.

Denote the input of the attention module as $X_Q \in \mathbb{R}^{N_q \times D}$, $X_K, X_V \in \mathbb{R}^{N_k \times D}$. N_q and N_k are the number of query tokens and key tokens, respectively. Note that, for simplicity, the batch and head dimensions are omitted. The attention score matrix is calculated as follows:

$$A_s = X_Q W_Q \cdot (X_K W_K)^T / \sqrt{d_h} \in \mathbb{R}^{N_q \times N_k}, \tag{4}$$

where $W_Q, W_K \in \mathbb{R}^{D \times d_h}$ are weights matrix of QK linear layer, d_h is the dimension for each head. The attention scores are normalized into attention weight matrix through the Softmax function as follows:

$$A_w = \text{Softmax}(A_s) = \left(\alpha_1, \dots, \alpha_{N_q}\right)^T \in \mathbb{R}^{N_q \times N_k}, \tag{5}$$

where attention weight vector $\alpha_i = (\alpha_{i,1}, \dots, \alpha_{i,N_k})^T \in \mathbb{R}^{N_k}$ represent the similarity of i-th query vector $q_i \in \mathbb{R}^{d_h}$ to all key vectors $\left\{k_j \in \mathbb{R}^{d_h}, j = 1, \dots, N_k\right\}$. A larger weight $\alpha_{i,j}$ indicates that q_i more closely matched to k_j, and the model considers the corresponding value vector v_j to be more important for the task at hand.

Given the above attention scores, we distinguish salient regions and define them as the ***Attention Focus***. A threshold factor $\theta \in (0, 1)$ is used to filter the significant values, resulting in a binarized ***Attention Focus Mask*** as follows:

$$M_A = M(A_w) = \mathbf{1}\left\{A_w > \theta \cdot \max\left(A_w\right)\right\} \in \mathbb{R}^{N_q \times N_k}, \tag{6}$$

where $\mathbf{1}\{\cdot\}$ represents an indicator function that returns 1 if the condition is true, otherwise 0.

The Attention Focus Mask implies critical perceptual information modeled by this attention module. On this basis, we can maintain the consistency of focusing patterns with the FP model, and thus enabling alignment at the semantic level. Specifically, we define the ***Attention Focus Overlap*** metric, which calculate the overlap of the Attention Focus Mask before and after quantization as follows:

$$\text{IoU}_{\text{AF}}\left(A_w, \hat{A}_w\right) = \frac{\left|M_A \cap \hat{M}_A\right|}{\left|M_A \cup \hat{M}_A\right|}, \tag{7}$$

where the variables with a hat superscript ($\hat{A}_w / \hat{M}_A$) denote the scores/masks from the quantized model. This metric lies in the range (0,1), and the distance function for PCC is defined as:

$$\text{Dist}_{pcc} = 1 - \text{IoU}_{\text{AF}}\left(A_w, \hat{A}_w\right). \tag{8}$$

Then, we leverage the defined Attention Focus Overlap metric to determine the optimal clipping boundaries (i.e. x_{low} and x_{up}) for QK activations. Unlike conventional methods limited by distributional matching, our proposed clipping metric operates at a semantic level, enabling more functionally effective clipping decisions. As shown in Figure 3, while MSE-based calibration causes severe attention degradation, our PCC achieves closer results to the FP model, efficiently preserves the semantic capabilities of the quantized attention module.

3.3 Prompt-Aware Reconstruction

Insight. Quantization reconstruction methods typically learn the quantization parameters $\{s, z, \alpha\}$ by locally minimize the error of block response under the supervision of the FP model as follows:

$$\min_{s,z,\alpha} \left\| \hat{O}^l - O^l \right\|_2^2, \tag{9}$$

where $O^l, \hat{O}^l$ are the outputs of l-th block from the FP model and quantized model, respectively. α is the adaptive rounding factor, an extra weight quantization parameter introduced by AdaRound (Nagel et al. 2020). While this method effectively enhances PTQ performance on conventional models, it struggles with the SAM model due to severe overfitting to pure image information. Specifically, Designed to follow the user's prompt, SAM's image embeddings interacting with prompts embeddings in mask decoder. Therefore, simply considering local visual response overlooks the prompt's intent, potentially introducing redundant information that disrupts the image-prompt interaction, resulting in distorted segmentation results.

Interaction Response Reconstruction. To address the aforementioned issue, we incorporate the interaction between prompt and image tokens into the reconstruction process, enabling effective image-prompt matching under the supervision of the FP model. Specifically, instead of aligning the raw visual response, we reconstruct the hybrid image tokens that integrate the prompt information by interacting in the cross-attention of mask decoder.

As illustrated in Figure 4, we utilize the off-the-shelf cross attention module in SAM's mask decoder to incorporate prompt information for the image tokens, without additional components or extra training. In this way, the hybrid image tokens can be obtained as follows:

$$T_{ip}^k = \text{TwoWayTransformer}(T_i^k, T_p), \tag{10}$$

where T_p denotes prompt tokens encoded by prompt encoder, T_i^k denotes image tokens derived from the outputs of the k-th stage. TwoWayTransformer refers to the Two-Way Transformer module detached from the mask decoder, which is designed for efficient image-prompt interaction. Subsequently, these tokens are reconstructed to learn the quantization parameters, i.e., minimizing the L2 distance between the hybrid image tokens and the corresponding FP response as follows:

$$\min_{s,z,\alpha} \left\| \hat{T}_{ip}^k - T_{ip}^k \right\|_2^2, \tag{11}$$

where T_{ip}^k and $\hat{T}_{ip}^k$ are tokens from the FP model and the quantized model, respectively.

Layer Skipping Interaction. To boost the efficiency of reconstruction, we customize two designs based on SAM's characteristics. First, we partition the Transformer layers into multiple stages and adopt stage-wise learning, where the quantization parameters within each stage are jointly optimized. Previous work (Shabanovi et al. 2024) has demonstrated the superiority of jointly optimizing multiple quantization blocks in reconstruction, as it effectively captures weight correlations across blocks. Building on this insight and SAM's architecture, we define global attention layers as boundaries for stage partitioning. For example, in SAM-B, layers $L2, L5, L8$ and $L11$ employ global attention, while the remaining layers utilize window attention. Accordingly, layers $\{L0, L1, L2\}$ constitute stage 0, and this pattern repeats for subsequent stages.

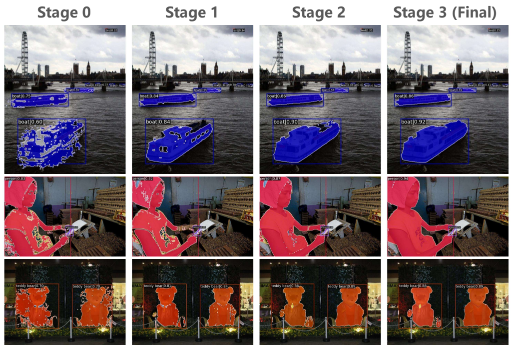


Figure 5: Segmentation results with image tokens from different stages. The output features of each stage are capable of skipping subsequent propagate while ensuring competent segmentation, with quality improving at deeper stages.

Furthermore, to promote efficient image-prompt interactions, the outputs of learning stage skip subsequent layers and directly pass the neck for interaction:

$$T_i^k = \text{Neck}((\textstyle\prod_{i=0}^{k} \text{Stage}^k)(E_i)), \tag{12}$$

where E_i denotes patch embeddings, Stage^k denotes the k-th stage, $\prod_{i=0}^{k} \text{Stage}^k$ represents the sequential composition of stages from stage 0 to stage k, Neck denotes SAM's Neck module for dimension reduction of image features. This layer-skipping strategy avoids prohibitively high computational cost of complete forward passes and gradient backpropagation, thus introducing slight additional overhead compared to traditional local response reconstruction. In addition to efficiency benefits, skipping deeper layers also brings performance advantages. With shortened optimization paths, it not only preserves local distribution properties, but also avoids the potential loss of interactive information due to long propagation processes.

To validate the feasibility of using these layer-skipping tokens directly for interaction, we tentatively use them as the final visual representation for mask prediction. The results shown in Figure 5 indicate that these immature tokens can produce rational segmentation results without further processing by deeper layers. This provides implicit evidence that the layer-skipping design is capable of effective interaction, with the outputs of each stage functioning as image embeddings at different levels of semantic granularity.

4 Experiments

4.1 Experimental Setup

Tasks and Datasets. Our experiments include three prevalent tasks with SAM: instance segmentation, oriented object detection, and semantic segmentation.

Detector	Faster R-CNN						YOLOX						H-Deformable-DETR						DINO					
Model	SAM-B		SAM-L		SAM-H		SAM-B		SAM-L		SAM-H		SAM-B		SAM-L		SAM-H		SAM-B		SAM-L		SAM-H	
FP	33.4		36.4		37.2		37.0		40.0		41.0		38.2		41.5		42.0		44.5		48.6		49.1	
Prec. (W/A)	6/6	4/4	6/6	4/4	6/6	4/4	6/6	4/4	6/6	4/4	6/6	4/4	6/6	4/4	6/6	4/4	6/6	4/4	6/6	4/4	6/6	4/4	6/6	4/4
MinMax	9.2	-	32.9	-	31.9	-	10.7	-	37.5	-	36.1	-	10.9	-	38.6	-	37.3	-	11.2	-	44.7	-	42.8	-
Percentile	10.9	-	33.5	-	32.0	-	12.0	-	38.0	-	36.3	-	12.3	-	39.0	-	37.5	-	14.0	-	45.4	-	43.1	-
OMSE	11.9	-	33.9	5.4	33.1	7.4	13.5	-	38.4	6.1	37.5	7.8	15.0	-	39.6	6.2	38.6	7.7	16.6	-	45.9	6.8	44.5	8.3
PTQ4SAM-S	15.4	-	35.7	18.1	36.0	24.1	17.4	-	**40.0**	20.6	**40.3**	26.7	17.9	-	41.0	20.9	**41.3**	27.3	20.4	-	47.7	23.1	48.1	30.5
SAQ-SAM*	**29.8**	**2.8**	**35.7**	**20.9**	**36.1**	**25.0**	**34.0**	**2.1**	**40.0**	**24.2**	40.2	**28.4**	**36.4**	**2.6**	**41.1**	**24.3**	**41.3**	**28.2**	**39.4**	**3.5**	**48.0**	**27.8**	**48.2**	**31.6**
AdaRound	23.1	-	34.3	8.7	33.7	14.5	26.4	-	38.9	11.1	38.3	16.7	27.2	-	39.9	8.0	39.4	16.3	31.2	1.2	46.6	8.8	46.0	18.2
BRECQ	24.1	-	34.2	10.7	33.7	15.1	26.1	-	38.9	12.0	38.3	16.3	27.9	-	39.9	11.1	39.5	15.5	31.8	3.6	46.6	12.3	46.0	17.6
QDrop	29.3	13.0	35.2	22.6	36.3	32.3	33.6	13.3	39.7	25.3	40.4	35.8	34.3	13.2	40.5	25.8	41.4	36.5	38.9	11.2	47.5	27.5	48.3	41.7
PTQ4SAM-L	30.3	16.0	35.8	28.7	36.5	33.5	34.3	18.4	40.3	31.6	40.7	37.6	35.1	17.3	41.2	32.1	41.6	38.4	40.4	14.4	**48.3**	36.6	48.7	43.9
SAQ-SAM	**32.0**	**27.7**	**35.9**	**34.5**	**36.6**	**35.9**	**35.9**	**30.3**	**40.3**	**39.0**	**40.9**	**39.9**	**37.1**	**31.8**	**41.3**	**39.8**	**41.8**	**40.7**	**42.4**	**33.8**	**48.3**	**46.3**	**48.9**	**47.4**

Table 1: Quantization results of instance segmentation on COCO dataset with various detectors providing prompts. We abbreviate 'Precision' as 'Prec.', 'W/A'=x/y denotes the weights and activations are quantized to x bit and y bit, respectively, and '-' indicates the mAP is below 1. Rows with green background indicate statistic-based methods, while blue background indicate learning-based methods. The bolded values represent the best-performing results under the same settings.

Method	SAM-B			SAM-L			SAM-H		
	FP	W6A6	W4A4	FP	W6A6	W4A4	FP	W6A6	W4A4
Percentile	63.93	10.35	2.62	64.35	60.98	10.06	64.49	60.08	14.55
OMSE		18.89	1.11		61.55	23.50		61.11	17.49
PTQ4SAM-S		19.90	0.58		60.88	10.25		62.99	40.14
SAQ-SAM*		**51.84**	**4.63**		**62.70**	**27.83**		**63.35**	**41.63**
AdaRound	63.93	34.05	-	64.35	63.44	23.18	64.49	62.73	24.45
BRECQ		34.40	-		63.60	26.89		62.58	25.98
QDrop		59.27	41.96		63.86	50.11		62.83	55.87
PTQ4SAM-L		60.33	44.18		63.91	56.29		64.36	56.01
SAQ-SAM		**60.35**	**44.39**		**64.38**	**58.03**		**64.58**	**60.17**

Table 2: Results of oriented object detection on DOTA dataset. Our SAQ-SAM shows consistent superiority in both comparison groups.

- In instance segmentation, various object detectors provide prompting boxes for SAM to obtain segmentation masks, including Faster R-CNN (Ren et al. 2015), YOLOX (Ge et al. 2021), H-Deformable-DETR (Jia et al. 2023), and DINO (Zhang et al. 2022), with MS-COCO (Lin et al. 2014) as evaluation dataset.
- In oriented object detection, with a detector generating horizontal boxes as prompts, SAM predicts masks that are subsequently transformed into resulting rotated boxes. We do experiment on DOTA (Ding et al. 2021) dataset with FCOS (Tian et al. 2019) detector.
- In the semantic segmentation, SAM is employed to enhance the quality of segmentation masks produced by traditional semantic segmentor. We employ SegFormer (Xie et al. 2021) as the semantic branch and evaluate on ADE20K (Zhou et al. 2017) dataset.

Implementation Details. We maintain consistant experimental setting with baselines (Lv et al. 2024) for fair comparision. Specifically, the quantization scheme employs per-tensor asymmetric quantization for activations and per-channel asymmetric quantization for weights. The weights are calibrated using the MSE. The activation calibration set contains 32 images randomly sampled from the training dataset. For the PCC, we set threshold factor θ to 0.5. All QK activations (i.e., the inputs and outputs of Query Linear and Key Linear) are calibrated based on Attention Focus Overlap metric using the first sample. For PAR, the image encoder employs per-stage learning, while the mask decoder employs per-layer learning with 2000 iterations. Exceptionally, the additional final cross-attention block learns 10000 iterations due to its large loss. In general, compared to PTQ4SAM settings, which carry per-block learning with 20000 iterations, our method requires much lower time cost.

Baseline Methods. We take advanced PTQ4SAM as baseline, because both are dedicated to refining mask decoder quantization. For comprehensive comparison, we follow the PTQ4SAM settings and divide the methods into two groups according to whether there is a learning process. For statistic-based PTQ methods, we integrated PCC into PTQ4SAM-S, which is referred to as SAQ-SAM*. And for learning-based PTQ, we build on PTQ4SAM-L and modified the reconstruction method from QDrop (Wei et al. 2022) to our PAR, with PCC as pre-calibration process.

4.2 Performance Evaluation

Instance Segmentation. In the task of instance segmentation, as shown in Table 1, our method consistently shows superior performance. Compared to baseline PTQ4SAM-S, SAQ-SAM* achieves a nearly twofold improvement for 6-bit SAM-B, For SAM-L and SAM-H, the advantages are particularly evident in low-bit settings. For example, SAQ-SAM* increases the mAP of 4-bit SAM-L from 23.1% to 27.8% with DINO. For reconstruction methods, compared to PTQ4SAM-L, our methods achieve significant improvement

Method	SAM-B			SAM-L		
	FP	W6A6	W4A4	FP	W6A6	W4A4
PTQ4SAM-S	33.15	31.16	31.08	33.61	33.48	22.64
SAQ-SAM*		**32.90**	**31.29**		**33.55**	**25.64**
AdaRound	33.15	32.34	31.78	33.61	32.99	31.97
BRECQ		32.27	31.78		33.04	31.98
QDrop		32.57	31.79		33.58	32.67
PTQ4SAM-L		32.65	31.85		**33.66**	32.82
SAQ-SAM		**33.04**	**32.53**		33.63	**33.30**

Table 3: Results of semantic segmentation on ADE20K dataset. Our SAQ-SAM demonstrate effective in improving segmentation performance.

with lower computational costs. For example, SAQ-SAM improves the 4-bit SAM-B by 14.5% mAP and achieves near-lossless accuracy for 4-bit SAM-L and SAM-H.

Oriented Object Detection. In the oriented object detection task, as shown in Table 2, our method also yields excellent performance, demonstrating generality for accurate segmentation and orientation of small targets. For example, in statistic-based setting, our SAQ-SAM* boost 6-bit SAM-B performance to a usable level (from 19.9% to 51.84%). In learning-based setting, SAQ-SAM improves the mAP of 4-bit SAM-L from 56.29% to 58.03%.

Semantic Segmentation. In the task of semantic segmentation, the accuracy degradation caused by quantization is not significant due to the sophisticate post-processing steps. Nevertheless, as shown in Table 3, our methods also show superiority. For example, SAQ-SAM* increases the accuracy of 6-bit SAM-B by 1.74% mIoU, even surpassing the learning-based PTQ4SAM-L. And SAQ-SAM further boosts the performance close to FP.

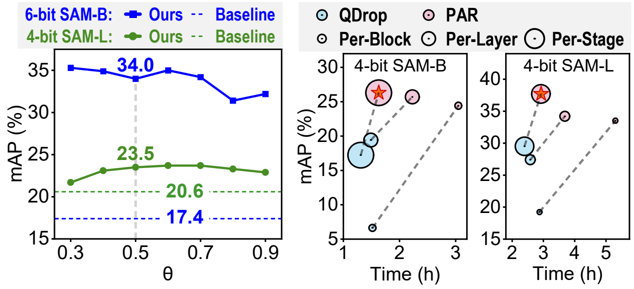
4.3 Ablation Study

With PTQ4SAM-L as baseline, we explore the effectiveness of SAQ-SAM components. As shown in Table 4, while PAR outperforms the baseline, when PCC act as pre-calibration for PAR, the combination of both achieves the best performance.

Detector	Method		SAM-B		SAM-L		SAM-H	
	PAR	PCC	W6A6	W4A4	W6A6	W4A4	W6A6	W4A4
YOLOX	✗	✗	34.3	18.4	**40.3**	31.6	40.7	37.6
	✓	✗	35.7	26.2	40.2	38.9	40.8	39.4
	✓	✓	**35.9**	**30.3**	**40.3**	**39.0**	**40.9**	**39.9**
DINO	✗	✗	40.4	14.4	**48.3**	36.6	48.7	43.9
	✓	✗	42.3	30.2	**48.3**	46.1	48.8	**47.4**
	✓	✓	**42.4**	**33.8**	**48.3**	**46.3**	**48.9**	**47.4**

Table 4: Ablation study of SAQ-SAM components.

Attention Focus Threshold Setting of PCC. In the proposed PCC module, we use the threshold factor θ to define the salient region of attention, which is typically set at 0.5. To explore the impact of this hyperparameter setting, we perform a grid search on it. As shown in Figure 6 (a), with changes in θ, the performance of PCC consistently outperforms the baseline method, albeit with slight fluctuations, where 0.5 is a modest choice. Notably, our method demonstrates exceptional robustness to the setting of θ.



(a) Grid search for θ of PCC (b) Efficiency study of PAR

Figure 6: Ablation study of PCC and PAR. Figure (a) shows PCC is robust to the setting of θ. Figure (b) shows per-stage PAR achieves optimal efficiency-accuracy trade-off.

Efficiency-Accuracy Study of PAR. To further verify the superiority of PAR, we explore its time consumption and performance with different granularity of optimization units, and compare it with QDrop-based reconstruction. As shown in Figure 6, per-stage scheme reduces the time cost of parameter learning, as well as boosting reconstruction performance, where per-stage PAR obtains optimal efficiency-accuracy trade-off. Furthermore, PAR performs better than QDrop under all granularity levels, highlighting the contribution of image-prompt interactions.

PCC could potentially serve as an orthogonal technique for combining with other advanced quantization methods. Since our method does not introduce any additional model components, the compression and acceleration benefits of quantized SAM are consistent with that reported in PTQ4SAM paper.

5 Conclusion

In this paper, we propose SAQ-SAM, a post-training quantization framework for the SAM model. To address the issue of extreme outliers in mask decoder's QK activations, we introduce a semantically-preserved outlier clipping approach, which can effectively suppress significant outliers without smoothing or isolation. Specifically, we determine the optimal clipping boundary by minimizing the attention focus deviation caused by quantization, overcoming the limitations of distribution-based methods. Furthermore, we incorporate image-prompt interactions into reconstruction, which learns the correlation between visual features and prompt intent, thereby achieving alignment at both semantic and distributional levels. To improve the interaction efficiency, we also introduce a layer-skipping strategy and a stage-partitioning approach. Extensive experiments show that our method significantly outperforms baseline methods, especially in low-bit scenarios. In the future, one could extend this approach to accelerate SAM2, which is a promising method for video segmentation acceleration.

Acknowledgments

This work was supported in part by the Strategic Priority Research Program of Chinese Academy of Sciences under Grant Number XDB1100000; in part by the National Natural Science Foundation of China under Grant Number 62276255; in part by the Postdoctoral Fellowship Program of CPSF under Grant Number GZC20251175; in part by the National Key Research and Development Program of China under Grant Number 2022ZD0119402; in part by the Beijing Natural Science Foundation Haidian Original Innovation Joint Fund Project under Grant Number L232035, L222154.

References

Chen, J.; Yang, Z.; and Zhang, L. 2023. Semantic Segment Anything. https://github.com/fudan-zvg/Semantic-Segment-Anything.

Chen, K.; Liu, C.; Chen, H.; Zhang, H.; Li, W.; Zou, Z.; and Shi, Z. 2024. RSPrompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–17.

Chen, Z.; Fang, G.; Ma, X.; and Wang, X. 2023. SlimSAM: 0.1% Data Makes Segment Anything Slim. *arXiv preprint arXiv:2312.05284*.

Cheng, J.; Ye, J.; Deng, Z.; Chen, J.; Li, T.; Wang, H.; Su, Y.; Huang, Z.; Chen, J.; Jiang, L.; et al. 2023. Sam-med2d. *arXiv preprint arXiv:2308.16184*.

Choi, J.; Wang, Z.; Venkataramani, S.; Chuang, P. I.-J.; Srinivasan, V.; and Gopalakrishnan, K. 2018. Pact: Parameterized clipping activation for quantized neural networks. *arXiv preprint arXiv:1805.06085*.

Ding, J.; Xue, N.; Xia, G.-S.; Bai, X.; Yang, W.; Yang, M.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; and Zhang, L. 2021. Object Detection in Aerial Images: A Large-Scale Benchmark and Challenges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–1.

Ge, Z.; Liu, S.; Wang, F.; Li, Z.; and Sun, J. 2021. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*.

Gholami, A.; Kim, S.; Dong, Z.; Yao, Z.; Mahoney, M. W.; and Keutzer, K. 2021. A survey of quantization methods for efficient neural network inference. *arXiv preprint arXiv:2103.13630*.

Jia, D.; Yuan, Y.; He, H.; Wu, X.; Yu, H.; Lin, W.; Sun, L.; Zhang, C.; and Hu, H. 2023. Detrs with hybrid matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19702–19712.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.

Li, C.; Qi, L.; and Geng, X. 2025. A SAM-guided two-stream lightweight model for anomaly detection. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21(2): 1–23.

Li, Y.; Gong, R.; Tan, X.; Yang, Y.; Hu, P.; Zhang, Q.; Yu, F.; Wang, W.; and Gu, S. 2021. Brecq: Pushing the limit of post-training quantization by block reconstruction. *arXiv preprint arXiv:2102.05426*.

Li, Y.; Mao, H.; Girshick, R.; and He, K. 2022a. Exploring plain vision transformer backbones for object detection. In *European conference on computer vision*, 280–296. Springer.

Li, Z.; Chen, M.; Xiao, J.; and Gu, Q. 2023a. Psaq-vit v2: Toward accurate and general data-free quantization for vision transformers. *IEEE Transactions on Neural Networks and Learning Systems*.

Li, Z.; and Gu, Q. 2023. I-vit: Integer-only quantization for efficient vision transformer inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 17065–17075.

Li, Z.; Liu, X.; Zhang, J.; and Gu, Q. 2024a. Repquant: Towards accurate post-training quantization of large transformer models via scale reparameterization. *arXiv preprint arXiv:2402.05628*.

Li, Z.; Liu, X.; Zhu, B.; Dong, Z.; Gu, Q.; and Keutzer, K. 2023b. Qft: Quantized full-parameter tuning of llms with affordable resources. *arXiv preprint arXiv:2310.07147*.

Li, Z.; Long, X.; Xiao, J.; and Gu, Q. 2024b. HTQ: Exploring the High-Dimensional Trade-Off of mixed-precision quantization. *Pattern Recognition*, 156: 110788.

Li, Z.; Ma, L.; Chen, M.; Xiao, J.; and Gu, Q. 2022b. Patch similarity aware data-free quantization for vision transformers. In *European conference on computer vision*, 154–170. Springer.

Li, Z.; Ma, L.; Long, X.; Xiao, J.; and Gu, Q. 2022c. Dual-discriminator adversarial framework for data-free quantization. *Neurocomputing*, 511: 67–77.

Li, Z.; Xiao, J.; Yang, L.; and Gu, Q. 2023c. Repq-vit: Scale reparameterization for post-training quantization of vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 17227–17236.

Li, Z.; Zhang, J.; and Gu, Q. 2024. Privacy-Preserving SAM Quantization for Efficient Edge Intelligence in Healthcare. *arXiv preprint arXiv:2410.01813*.

Lin, J.; Tang, J.; Tang, H.; Yang, S.; Chen, W.-M.; Wang, W.-C.; Xiao, G.; Dang, X.; Gan, C.; and Han, S. 2024. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of Machine Learning and Systems*, 6: 87–100.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, 740–755. Springer.

Liu, J.; Niu, L.; Yuan, Z.; Yang, D.; Wang, X.; and Liu, W. 2023. Pd-quant: Post-training quantization based on prediction difference metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24427–24437.

Liu, X.; Ding, X.; Yu, L.; Xi, Y.; Li, W.; Tu, Z.; Hu, J.; Chen, H.; Yin, B.; and Xiong, Z. 2024a. PQ-SAM: Post-training Quantization for Segment Anything Model. In *European Conference on Computer Vision*, 420–437. Springer.

Liu, X.; Li, Z.; and Gu, Q. 2024. Dilatequant: Accurate and efficient diffusion quantization via weight dilation. *arXiv preprint arXiv:2409.14307*.

Liu, X.; Li, Z.; and Gu, Q. 2025. Cachequant: Comprehensively accelerated diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 23269–23280.

Liu, X.; Li, Z.; Xiao, J.; and Gu, Q. 2024b. EDA-DM: Enhanced Distribution Alignment for Post-Training Quantization of Diffusion Models. *arXiv preprint arXiv:2401.04585*.

Liu, Z.; Wang, Y.; Han, K.; Zhang, W.; Ma, S.; and Gao, W. 2021. Post-training quantization for vision transformer. *Advances in Neural Information Processing Systems*, 34: 28092–28103.

Lv, C.; Chen, H.; Guo, J.; Ding, Y.; and Liu, X. 2024. Ptq4sam: Post-training quantization for segment anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15941–15951.

Ma, J.; He, Y.; Li, F.; Han, L.; You, C.; and Wang, B. 2024. Segment anything in medical images. *Nature Communications*, 15(1): 654.

Nagel, M.; Amjad, R. A.; Van Baalen, M.; Louizos, C.; and Blankevoort, T. 2020. Up or down? adaptive rounding for post-training quantization. In *International conference on machine learning*, 7197–7206. PMLR.

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; et al. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Shabanovi, K.; Wiest, L.; Golkov, V.; Cremers, D.; and Pfeil, T. 2024. Interactions Across Blocks in Post-Training Quantization of Large Language Models. *arXiv preprint arXiv:2411.03934*.

Shao, W.; Chen, M.; Zhang, Z.; Xu, P.; Zhao, L.; Li, Z.; Zhang, K.; Gao, P.; Qiao, Y.; and Luo, P. 2023. Omniquant: Omnidirectionally calibrated quantization for large language models. *arXiv preprint arXiv:2308.13137*.

Tian, Z.; Shen, C.; Chen, H.; and He, T. 2019. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9627–9636.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wei, X.; Gong, R.; Li, Y.; Liu, X.; and Yu, F. 2022. Qdrop: Randomly dropping quantization for extremely low-bit post-training quantization. *arXiv preprint arXiv:2203.05740*.

Xiao, G.; Lin, J.; Seznec, M.; Wu, H.; Demouth, J.; and Han, S. 2023. Smoothquant: Accurate and efficient post-training quantization for large language models. In *International Conference on Machine Learning*, 38087–38099. PMLR.

Xiao, J.; Li, Z.; Li, J.; Yang, L.; and Gu, Q. 2024. Binaryvit: Towards efficient and accurate binary vision transformers. *IEEE Transactions on Circuits and Systems for Video Technology*.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34: 12077–12090.

Xiong, Y.; Varadarajan, B.; Wu, L.; Xiang, X.; Xiao, F.; Zhu, C.; Dai, X.; Wang, D.; Sun, F.; Iandola, F.; et al. 2024. Efficientsam: Leveraged masked image pretraining for efficient segment anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16111–16121.

Yu, Y.; Yang, X.; Li, Q.; Zhou, Y.; Da, F.; and Yan, J. 2023. H2RBox-v2: Incorporating symmetry for boosting horizontal box supervised oriented object detection. *Advances in Neural Information Processing Systems*, 36: 59137–59150.

Zhang, C.; Han, D.; Qiao, Y.; Kim, J. U.; Bae, S.-H.; Lee, S.; and Hong, C. S. 2023. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*.

Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L. M.; and Shum, H.-Y. 2022. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*.

Zhang, Z.; Cai, H.; and Han, S. 2024. Efficientvit-sam: Accelerated segment anything model without performance loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7859–7863.

Zhao, X.; Ding, W.; An, Y.; Du, Y.; Yu, T.; Li, M.; Tang, M.; and Wang, J. 2023. Fast segment anything. *arXiv preprint arXiv:2306.12156*.

Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; and Torralba, A. 2017. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 633–641.