Mix-QSAM2: Mixed-Precision Quantization for High Fidelity Segmentation in Resource Constrained Scenarios

Yuzhe Duan[1*], Xuanxuan Ren[2*], Guizhe Dong[3], Xu Yang[3†], Yanhua Yang[1†],

[1]School of Computer Science and Technology, Xidian University, Xi'an 710071, China
[2]School of Artificial Intelligence, Xidian University, Xi'an 710071, China
[3]School of Electronic Engineering, Xidian University, Xi'an 710071, China
1.yzduan@gmail.com, {xxren, 24021211684}@stu.xidian.edu.cn, xuyang.xd@gmail.com, yanhyang@xidian.edu.cn

Abstract

The Segment Anything Model 2 (SAM2) has established a new benchmark for high-precision image and video segmentation, offering significant potential for a wide range of computer vision tasks. Despite its impressive performance, the model's substantial computational and memory requirements present a significant obstacle to its practical deployment on resource-constrained devices. In this paper, we introduce a novel framework for optimizing SAM2 through two synergistic, importance-driven strategies: quantization and memory management. Specifically, an Importance-driven Mixed-Precision Quantization scheme, which analyzes the sensitivity of each layer using a Weight-Activation Importance Score, is employed to enable a targeted bit-width assignment, preserving model accuracy by keeping critical layers at higher precision. Then, the Selective Importance-driven Synthesis (SIS) mechanism is proposed to address the inefficient accumulation of redundant data in the memory bank. SIS intelligently compresses the memory by identifying the most contextually similar historical frames and synthesizing them into a single, representative feature, thereby preserving informational diversity while enhancing temporal context understanding. Extensive experiments on the COCO and SA-V benchmarks validate our approach, showing that our optimized model consistently outperforms state-of-the-art quantization methods. Our work provides a principled framework for the co-design of quantization and dynamic memory management, offering a practical path toward deploying powerful video segmentation models in real-world applications.

Introduction

Large-scale foundation models have largely propelled recent advancements in visual segmentation, with SAM2 (Ravi et al. 2024) representing a significant milestone in this evolution. These models demonstrate remarkable performance across diverse tasks due to their extensive pre-training and sophisticated architectures. However, the substantial computational demands of these models present significant challenges for real-world deployment, particularly in resource-constrained environments with limited memory and processing power. For segmentation tasks, these limitations pose significant hurdles, as achieving both fine-grained spatial preservation and high computational efficiency is paramount for real-world deployment.

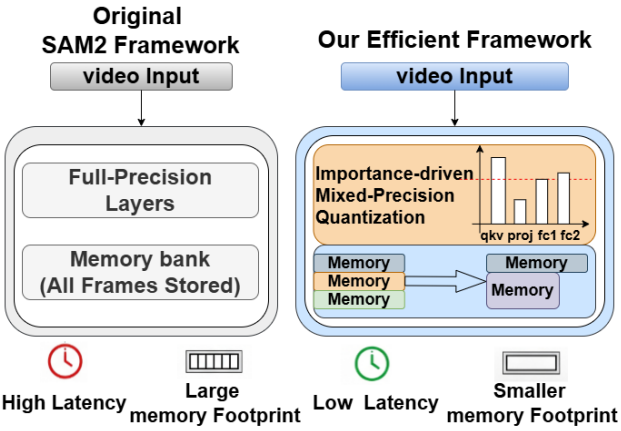

Figure 1: A comparison of our proposed efficient framework (right) against the original SAM2 (left). Our method combines Importance-driven Mixed-Precision Quantization and Selective Importance-driven Synthesis to address the original model's high latency and memory footprint while maintaining high performance.

To address these challenges, significant efforts have been made to adapt large models for edge deployment. MobileSAM (Zhang et al. 2023) reduces the encoder's computational load with a lightweight TinyViT architecture (Khan et al. 2022; Wu et al. 2022), FastSAM (Zhao et al. 2023) reframes the problem for an efficient object detector, and TinySAM (Shu et al. 2025) employs knowledge distillation to train a smaller model. Another prominent approach is quantization (Nagel et al. 2021), which converts a model's floating-point parameters into low-bit representations. As a practical approach, Post-Training Quantization (PTQ) (Frantar et al. 2022; Li et al. 2023; Dettmers et al. 2023; Huang et al. 2024b) achieves this by calibrating the network with a small, unlabeled dataset, thus avoiding costly retraining.PTQ4SAM (Lv et al. 2024)successfully applying it to the original SAM. Despite its promise, current quantization methods often rely on uniform bit-widths (Ren et al.

*Equal contribution
†Corresponding author

2025) or static mixed-precision schemes (Tang et al. 2022; Dong et al. 2019), overlooking the dynamic importance of different network components during inference.

In addition to model quantization, memory management, which is crucial for understanding temporal dynamics in videos, has undergone significant evolution (Zhang et al. 2025). Early video object segmentation methods (Hu, Huang, and Schwing 2018; Avinash Ramakanth and Venkatesh Babu 2014) often process the current frame along with most or all previous frames, leading to significantly increased inference overhead. More advanced models, such as SAM2, adopt a refined approach where the current frame interacts with a memory bank of representations from prior frames through a cross-attention mechanism (Vaswani et al. 2017). However, these techniques often rely on fixed caching strategies and simple heuristics, and are largely decoupled from the semantic content being processed. While techniques such as quantization and memory optimization (Rajbhandari et al. 2020)have shown promise in reducing model footprint, existing approaches, primarily focused on image processing, typically address these aspects independently, thereby missing opportunities for synergistic optimization.

This paper takes the preservation of key information in the large model reasoning process as the starting point and constructs a SAM2 acceleration optimization scheme with importance evaluation as the core. Our key information research focuses on two aspects: parameter importance and content importance, comprising two fundamental innovations in a unified framework. Specifically, our Importance-driven Mixed-Precision Quantization strategy operates at the layer level, utilizing a post-training, calibration-based approach. It computes a Weight-Activation Importance Score for each layer based on the magnitudes of weights and activation statistics, which then dynamically guides bit-width assignment to minimize accuracy loss within a given computational budget. Second, our Selective Importance-driven Synthesis (SIS) mechanism introduces a content-aware approach to memory compression. Instead of relying on fixed temporal rules, SIS dynamically evaluates all historical frames based on their content similarity to the present context and synthesizes only the most redundant subset, efficiently reducing memory overhead while preserving the diversity of the long-term temporal context.

Our experiments demonstrate that the proposed framework effectively incorporates quantization and memory optimizations, while maintaining a competitive level of performance on standard segmentation benchmarks. This is achieved without any architectural modifications, confirming our approach as a practical enhancement.

The remainder of this paper is structured as follows: Section 2 surveys related work on network quantization and efficient vision models. Section 3 outlines our proposed framework, which includes hierarchical quantization and importance-driven memory management schemes. Section 4 presents and analyzes our experimental results, followed by a discussion of their implications. Finally, Section 5 provides concluding remarks and suggests avenues for future research.

Relation Works

Model Quantization

The emergence of large foundation models, such as the Segment Anything Model (Kirillov et al. 2023), has highlighted a critical challenge: their substantial memory and computational demands hinder deployment on resource-constrained devices. Model quantization (Nagel et al. 2021) addresses this by converting high-precision floating-point parameters to low-bit integers. While Quantization-Aware Training (Lee, Kim, and Ham 2021; Li et al. 2022; Zhu, He, and Wu 2023) can achieve high accuracy, its requirement for full-dataset retraining is often prohibitively expensive. Consequently, Post-Training Quantization (Frantar et al. 2022; Li et al. 2023; Dettmers et al. 2023; Huang et al. 2024b) has become a more practical and widely researched alternative, as it only needs a small, unlabeled calibration set and avoids costly retraining. Early PTQ methods were primarily statistic-based, determining quantization parameters like the scaling factors and zero-point z by analyzing tensor distributions. Examples include MinMax (Jacob et al. 2018), which uses the absolute range, Percentile (Wu et al. 2020), which clips outliers, and OMSE (Choukroun et al. 2019), which minimizes the mean squared error. However, these methods are often suboptimal as they are susceptible to outlier values that are common in large models. This led to the development of more advanced learning-based PTQ methods, which reframe quantization as an optimization problem.

Instead of minimizing direct numerical error, these methods aim to reduce the reconstruction error of a layer's or a block's output, which is a better proxy for task performance. Seminal works in this area include AdaRound (Nagel et al. 2020), which introduced a learnable task-aware rounding mechanism instead of the standard nearest-rounding policy. BRECQ (Li et al. 2021) extended this by proposing block-wise reconstruction to handle inter-layer dependencies and using Fisher information to protect essential weights. QDrop (Wei et al. 2022) further improved robustness by introducing a Dropout-inspired regularization technique that makes the model more resilient to quantization noise. More recently, methods like PD-Quant (Liu et al. 2023) have sought to use more global information, such as the difference in the final model prediction, to guide the optimization. Furthermore, the rise of Transformer-based architectures has revealed that generic PTQs, such as Bimodal Integration (BIG), instance, PTQ4SAM (Lv et al. 2024), identified unique challenges in SAM, such as a "bimodal distribution" in specific activations that breaks standard quantizers. To address this, they proposed architecture-specific solutions like Bimodal Integration (BIG) to merge the distribution peaks and Adaptive Granularity Quantization (AGQ) to handle diverse post-Softmax distributions, demonstrating the need for co-designing quantization strategies with the model architecture.

Video Object Segmentation

The field of Video Object Segmentation (Ding et al. 2023; Yang, Wei, and Yang 2021; Zhao et al. 2025) is undergoing a paradigm shift from traditional appearance and motion-

based feature matching towards higher-level conceptual understanding, particularly for handling complex scenarios like occlusions (Perazzi et al. 2016). The core of this trend, often catalyzed by models like LVLMs (Radford et al. 2021), is to build a robust semantic representation of a target object over time, shifting the task from pure feature matching (Huang et al. 2024a) to concept reasoning (Barbiero et al. 2023). This evolution underscores the importance of an efficient and intelligent temporal memory mechanism that can preserve vital long-term context. It directly motivates our Selective Importance-driven Synthesis (SIS) mechanism for optimizing the SAM2 memory bank.

Methods

To achieve efficient and high-fidelity video object segmentation, we holistically optimize the SAM2 model's workflow. The baseline SAM2 pipeline processes videos in a streaming, frame-by-frame manner: an image encoder processes the current frame, its features are cross-attended with historical memory representations, and a mask decoder generates the final prediction. While effective, this workflow presents two key bottlenecks: the high computational cost of the full-precision encoder and the redundant accumulation of temporal information in its static memory bank. To address these challenges, we introduce a unified framework that optimizes this pipeline through two synergistic, importance-driven innovations. We propose an Importance-driven Mixed-Precision Quantization scheme to significantly compress the model and accelerate inference without substantial accuracy loss, coupled with a novel Selective Importance-driven Synthesis (SIS) mechanism that intelligently manages the temporal memory bank to enhance long-term tracking robustness. Together, these components transform the standard SAM2 pipeline into a resource-efficient and high-performance framework for video segmentation. The following sections will outline our enhancements built upon this foundation.

Importance-driven Mixed-Precision Quantization

Traditional PTQ is an effective model compression technique that maps high-precision floating-point weights (FP32) and activations to low-bit integers (e.g., INT8, INT4) without retraining. Its core is affine quantization, where for a floating-point tensor x, the quantization process can be expressed as:

$$x_q = \text{clamp}\left(\text{round}\left(\frac{x}{S} + Z\right), 0, 2^b - 1\right). \quad (1)$$

Here, x_q denotes the quantized low-bit integer representation, *b* is the bit-width, while *S* and *Z* are the scale factor and zero-point, respectively, which are calculated based on the dynamic range **[min(x), max(x)]** of *x* observed on a calibration dataset. However, this traditional PTQ approach, which applies a uniform bit-width to all layers, overlooks the varying sensitivity of different layers to quantization errors. To address this, we propose a mixed-precision allocation strategy driven by layer importance.

Layer Importance Quantification. We posit that a layer's importance is determined by the overall impact of its weights on the model's output. To quantify this importance, we define a **Weight-Activation Importance Score (S_l)** for each linear layer *l*. The calculation process is as follows:

First, for the input activation tensor $X^{(l)} \in R^{N \times C_{\text{in}}}$ of layer *l* where *N* is the number of tokens, and C_{in} is the number of input channels we measure the importance of each input channel *j* by its average energy (L2 norm) across the calibration set, yielding an input channel importance vector $A^{(l)} \in R^{C_{\text{in}}}$.Each element $A_j^{(l)}$ is computed as:

$$A_j^{(l)} = \left(\sum_{n=1}^{N} \left(X_{n,j}^{(l)}\right)^2\right)^{\frac{1}{2}}, \quad \text{for } j = 1, \ldots, C_{\text{in}}. \quad (2)$$

Second, we construct an element-wise importance matrix $H^{(l)}$ with the same dimensions as the weight matrix $W^{(l)} \in R^{C_{\text{out}} \times C_{\text{in}}}$,where C_{out} is the number of output channels. Each element $H_{i,j}^{(l)}$ is the product of the absolute weight $|W_{i,j}^{(l)}|$ and its corresponding input channel importance $A_j^{(l)}$:

$$H_{i,j}^{(l)} = |W_{i,j}^{(l)}| A_j^{(l)}. \quad (3)$$

Finally, the definitive importance score S_l for the entire layer *l* is defined as the mean of all elements in the importance matrix $H^{(l)}$:

$$S_l = \frac{1}{C_{\text{out}} \cdot C_{\text{in}}} \sum_{i=1}^{C_{\text{out}}} \sum_{j=1}^{C_{\text{out}}} H_{i,j}^{(l)}. \quad (4)$$

A higher score S_l indicates that the layer is more critical to the model's function and thus more sensitive to quantization perturbations.

Our chosen metric is not merely a heuristic but a computationally efficient approximation of complex second-order methods like Optimal Brain Surgeon (OBS) (Hassibi, Stork, and Wolff 1993), which aim to minimize layer-wise reconstruction error (Sun et al. 2023). The derivation from the complex, Hessian-based metric to a simple, efficient form is shown below:

$$S_{ij} \xrightarrow{\lambda=0} \left[|\mathbf{W}|^2 / \text{diag}\left((\mathbf{X}^T\mathbf{X})^{-1}\right)\right]_{ij} \xrightarrow{\text{diagonal}} \left[|\mathbf{W}|^2 / \left(\text{diag}(\mathbf{X}^T\mathbf{X})\right)^{-1}\right]_{ij} = \left(|W_{ij}| \cdot \|\mathbf{X}_j\|_2\right)^2. \quad (5)$$

As demonstrated, the complex second-order metric, which depends on the inverse Hessian matrix $(X^T X)$, simplifies to the squared value of our chosen importance metric under a diagonal approximation. This connection provides a strong theoretical foundation for our method, linking it to the minimization of reconstruction error while avoiding the prohibitive computational cost of a full second-order calculation.

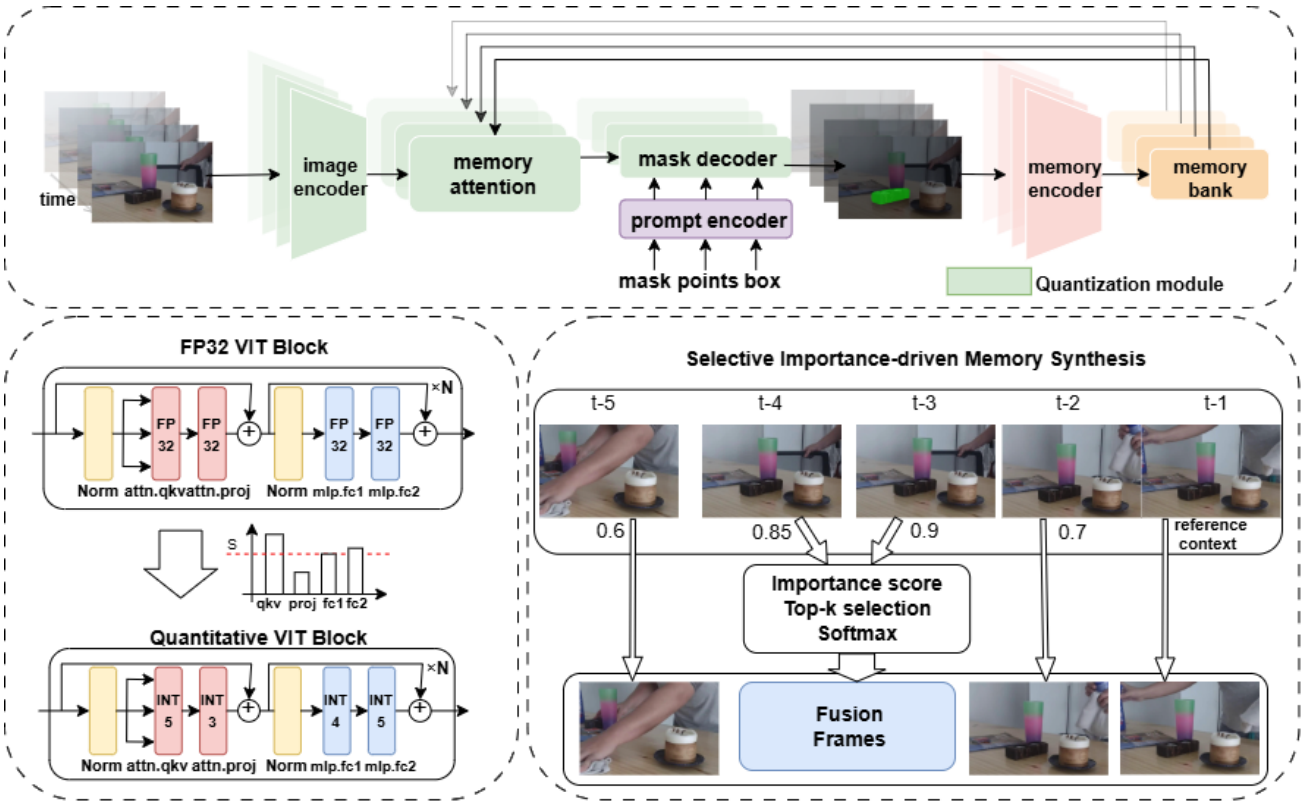

Figure 2: Our proposed optimization framework. The two core modules are: 1) Importance-driven Mixed-Precision Quantization (bottom-left), which assigns bit-widths based on layer importance to reduce computational costs while preserving performance. 2) Selective Importance-driven Memory Synthesis (bottom-right), which compresses temporal memory by identifying and fusing the most relevant historical frames.

Bit Allocation Strategy. After obtaining the importance scores for all layers, our goal is to assign an appropriate bit-width b_l to each layer l from a candidate set of precisions $\mathcal{B}$. This assignment aims to minimize the overall accuracy loss, weighted by importance, subject to a predefined target average bit-width B_{avg}. This can be formulated as a constrained optimization problem:

$$\min_{b_l \in \mathcal{B}} \sum_{l \in \mathcal{L}} S_l \cdot \mathcal{E}(b_l) \quad \text{subject to} \quad \frac{\sum_{l \in \mathcal{L}} b_l \cdot p_l}{\sum_{l \in \mathcal{L}} p_l} \leq B_{\text{avg}}, \tag{6}$$

where $\mathcal{L}$ is the set of all layers to be quantized, S_l is the importance score of layer l, and p_l is its number of parameters. $\mathcal{E}(b_l)$ represents the theoretical error introduced by quantizing a layer to b_l bits, which is a monotonically increasing function as bit-width decreases, such as $\mathcal{E}(b_l) \propto 2^{-b_l}$. The core idea of this formulation is to prioritize allocating more bits to layers with higher importance scores S_l to minimize their contribution to the total error.

Solving the combinatorial optimization problem directly is computationally intensive. Therefore, we design an efficient heuristic iterative adjustment algorithm to find an approximate optimal solution. The algorithm first performs an initial bit allocation based on the layer importance ranking, assigning higher initial precision to more essential layers. Subsequently, it iteratively fine-tunes the allocation to meet the average bit-width constraint. In each iteration, it calculates the current weighted average bit-width. If the value is higher than the target, the algorithm selects a layer with the most significant impact factor from those with higher precision and downgrades it to the next lower bit-width in the candidate set. Conversely, if the value is lower than the target, it selects the layer with the most significant number of parameters from those with lower precision and upgrades it to the next higher bit-width. This process is repeated until the target constraint is met, enabling the fast and efficient generation of a high-quality, importance-driven mixed-precision configuration.

Selective Importance-driven Synthesis

During long-term video tracking, the memory bank accumulates significant redundant information highly similar to recent frames. While temporal proximity is a basic heuristic for importance, it is content-agnostic and fails to distinguish between clear, high-quality memories and corrupted ones (e.g., from occluded or blurry frames). To overcome this, we propose a **Selective Importance-driven Synthesis (SIS)** mechanism, where every decision, from evaluation to information fusion, is explicitly driven by a quantitative,

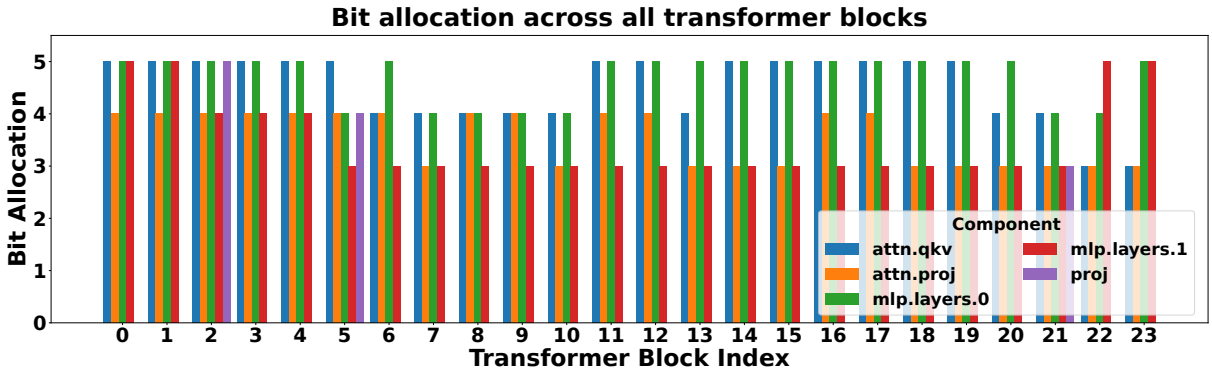

Figure 3: Mixed-precision bit allocation for the SAM-2B model. This figure illustrates the final bit width assigned to each layer. Candidate bit widths are 3, 4, 5.

content-based measure of importance.

The SIS mechanism is triggered when the memory bank exceeds its capacity. The process is initiated with Importance Quantification. We posit that the 'importance' of a historical frame lies in its contextual consistency and relevance to the immediate present. Therefore, we designate the feature representation of the most recent memory frame (from time `t-1`) as the reference context. The importance of every older historical frame is then quantified by calculating the cosine similarity between its feature representation and this reference. This similarity score provides a direct, content-based metric of each memory's redundancy and value. Following this quantification, all historical frames are subjected to Importance-based Ranking, creating a prioritized queue from the most important (most similar and redundant) to the least important (most unique).

This ranked queue of importance scores directly drives the core synthesis action. Unlike methods that perform global memory compression, SIS executes a targeted, importance-driven selective synthesis. The mechanism selects only a predefined number, `k`, of the highest-ranked historical frames (e.g., k=3)—those quantitatively identified as the most "important" and thus most redundant—for information synthesis. This selection is the first core decision driven by importance. The synthesis process is itself importance-driven: we employ a weighted averaging strategy to fuse these `k` frames, where the contribution of each frame is proportional to its importance score after normalization via a Softmax function. The weighted average is computed as:

$$F_{\text{synth}} = \sum_{i=1}^{k} w_i \cdot F_i, \quad \text{where} \quad w = \text{softmax}(\{s_1, \ldots, s_k\}), \tag{7}$$

where F_i and s_i are the feature and score of the i-th most important frame, and w_i is its corresponding normalized weight. This step ensures that the resulting synthesized feature, F_{synth}, is a robust representation whose characteristics are dominated by the most reliable and relevant historical information.

Finally, the Memory Bank is Reconstructed in an importance-aware manner. The new memory bank is a hybrid composition, consisting of: (1) the single, powerful synthesized feature that summarizes the most redundant context, F_{synth}; (2) all other historical frames that were not selected for fusion due to their lower importance scores (i.e., being more unique and dissimilar to the current context); and (3) the intact reference frame from `t-1`. The entire SIS process—from quantification and ranking to selective synthesis and reconstruction—thus constitutes a sophisticated, fully importance-driven loop. It ensures that memory compression is not a blind, uniform process, but a targeted intervention that intelligently preserves informational diversity while reducing redundancy, thereby significantly enhancing the model's long-term tracking robustness in complex scenarios.

Experiments

To comprehensively evaluate the effectiveness of our proposed Importance-driven Quantization and Synthesis Framework, this chapter presents a series of extensive experiments. We first introduce the benchmark datasets, evaluation protocols, and specific implementation details. Subsequently, we conduct in-depth quantitative and qualitative comparisons of our method against state-of-the-art PTQ methods on both image instance segmentation and promptable visual segmentation tasks. Finally, we systematically validate the individual contributions of each innovative component within our framework through rigorous ablation studies.

Experimental Setup

Datasets. All our experiments are conducted on two widely recognized public datasets. For the image instance segmentation task, we utilize the 2017 validation set of the MS-COCO dataset (Lin et al. 2014), which contains 5,000 images. For the video segmentation task, we perform evaluations on the SA-V dataset (Ravi et al. 2024), one of the largest video segmentation datasets to date, comprising over 50,000 real-world videos and 600,000 spatio-temporal masks.Tasks and Metrics. Our evaluation covers two core tasks. For Image Instance Segmentation, we adopt the standard mean Average Precision (mAP) as the primary evalu-

Method	Model-S		Model-B		Model-L	
	6-bit	4-bit	6-bit	4-bit	6-bit	4-bit
Full Precision (FP) mAP	*40.3*		*41.1*		*41.4*	
MinMax (Jacob et al. 2018)	10.9	-	35.5	-	35.1	-
Percentile (Wu et al. 2020)	12.2	-	36.0	-	35.4	-
OMSE (Choukroun et al. 2019)	13.3	-	36.4	5.9	36.5	7.6
AdaRound (Nagel et al. 2020)	26.2	-	37.8	10.6	36.8	12.7
BRECQ (Li et al. 2021)	25.9	-	37.8	12.0	36.7	12.3
Qdrop (Wei et al. 2022)	33.3	13.0	39.3	25.1	37.1	29.4
PTQ4SAM (Lv et al. 2024)	34.2	18.4	38.5	31.6	37.9	30.2
Ours	**38.8**	**34.3**	**40.3**	**34.2**	**38.9**	**32.0**

Table 1: Results of image instance segmentation on the COCO dataset (mAP). W6A6/W4A4 denote uniform quantization, while Avg-6/Avg-4 denote our mixed-precision method.

Method	Model-B			Model-L		
	FP	6-bit Setting	4-bit Setting	FP	6-bit Setting	4-bit Setting
Adaround (Nagel et al. 2020)	72.7	47.2 (W6A6)	25.6 (W4A4)	73.7	48.1 (W6A6)	27.3 (W4A4)
QDrop (Wei et al. 2022)	72.7	61.4 (W6A6)	33.4 (W4A4)	73.7	63.1 (W6A6)	35.8 (W4A4)
PTQ4SAM (Lv et al. 2024)	72.7	66.5 (W6A6)	40.8 (W4A4)	73.7	67.1 (W6A6)	40.4 (W4A4)
Ours	72.7	**68.7** (Avg-6)	**50.3** (Avg-4)	73.7	**68.6** (Avg-6)	**49.6** (Avg-4)

Table 2: Comparison of promptable visual segmentation performance (J&F) on the SA-V dataset. W6A6/W4A4 denote uniform quantization for competitor methods, while Avg-6/Avg-4 denote our mixed-precision scheme.

ation metric. For Promptable Visual Segmentation, we follow the common evaluation standard for this task, using the Jaccard Index $\mathcal{J}$ and the F-measure $\mathcal{F}$, and primarily report their average $\mathcal{J}\&\mathcal{F}$ as the comprehensive performance measure.

Implementation Details. Our experiments are based on three variants of the SAM2 model: Model-S, -B, and -L. We use all three models for the image segmentation task, and Model-B and -L for the video segmentation task. All comparison methods are evaluated under full-precision (FP) and two quantized configurations. The reference methods employ uniform 6-bit (W6A6) and 4-bit (W4A4) quantization. In contrast, our method uses a mixed-precision scheme, denoted as Avg-6 and Avg-4 for the respective target average bit-rates. For the calibration stage, we randomly selected 32 images and 8 videos. During model reconstruction, each module undergoes at least 20000 iterations. All experiments were conducted on a server equipped with four NVIDIA A6000 GPUs, running Ubuntu 20.04. The framework was implemented using PyTorch 2.6.0 and CUDA 12.4. Following conventional practice, the first and last layers of the network are excluded from quantization.

Image Instance Segmentation Results

The image instance segmentation results, averaged over three independent runs, are presented in Table 1. Our method demonstrates a consistent and significant performance advantage over all leading PTQ baselines across every model size and bit-rate tested. This advantage is particularly evident in 4-bit quantization settings, where our Avg-4 models not only lead the 4-bit category but also regularly surpass the performance of competitors' 6-bit quantized models.

Visual Segmentation Results

The results on the SA-V dataset further validate the robustness of our framework in processing temporal data. where all results are averaged over three separate runs, we compare our method against AdaRound, QDrop, and PTQ4SAM.

In the video task, the performance advantage of our method is even more pronounced, which we attribute to our SIS mechanism. Compared to the image task, video segmentation places higher demands on temporal understanding and memory management, and our SIS mechanism effectively enhances the model's ability in this regard by intelligently compressing redundancy while preserving diversity. This leads to strong results; for instance, on Model-B under the Avg-4 setting, our method achieves a $\mathcal{J}\&\mathcal{F}$ score of 50.3, leading all comparison methods. On a separate note, we observe that the quantized model's performance decreases when scaling from SAM2-Base to SAM2-Large. We speculate this is due to error accumulation, as errors from low-precision layers can compound in deeper networks and ultimately degrade performance.

Ablation Studies

To dissect the individual contributions of our two core innovations—Importance-driven Mixed-Precision Quantization (MP) and Selective Importance-driven Synthesis

Figure 4: Visual Segmentation Results

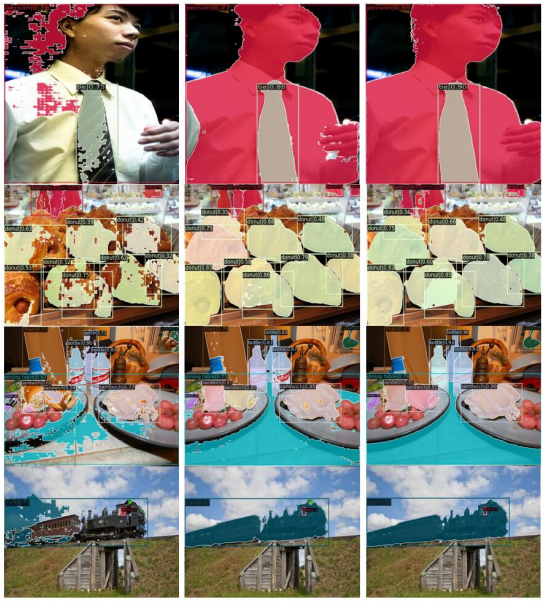

Figure 5: Result of image instance segmentation

Configuration	MP	SIS	$\mathcal{J}\&\mathcal{F}$ @ 4-bit	6-bit
Base (A)	×	×	40.8	66.5
+ MP (B)		×	49.2	67.0
+ SIS (C)	×		41.9	67.7
Ours (D)			**50.3**	**68.7**

Table 3: Ablation study on the SA-V dataset (Model-B). All results are $\mathcal{J}\&\mathcal{F}$ scores under the most aggressive 4-bit setting. The full-precision (FP) baseline score is 72.7. MP and SIS are abbreviations for our two proposed components. Δ denotes the improvement over the Base model.

(SIS)—we conducted a series of ablation studies on the SA-V dataset, Using the quantized SAM2 Model-B as the baseline. The results systematically validate our design choices.

As shown in Table 3, each of our proposed components makes a positive contribution to the final performance. Our MP scheme alone provides a significant performance boost, especially in the challenging 4-bit setting, validating the effectiveness of our importance-driven bit allocation.

Qualitative Results

Figure 4 illustrates the qualitative results for the promptable visual segmentation task. Our method demonstrates robust tracking capabilities in challenging scenarios, such as partial occlusions (row 1) and continuously moving objects (rows 2 and 3). Notably, our framework consistently generates fine-grained masks, accurately delineating detailed object contours. For image instance segmentation, Figure 5 provides a qualitative comparison. Existing methods often suffer from incomplete masks (e.g., the monitor in row 1) and imprecise object boundaries, particularly in complex scenes (rows 2 and 3). In contrast, our approach consistently generates masks with superior completeness and boundary detail across all these challenging cases.

Conclusion

In this paper, we propose an importance-driven framework to address the computational and memory bottlenecks of the large-scale video segmentation model, SAM2. Our framework uniquely integrates two synergistic mechanisms: an Importance-driven Mixed-Precision Quantization scheme for efficient, accuracy-preserving model compression, and a Selective Importance-driven Synthesis (SIS) mechanism to manage temporal memory and enhance long-term tracking robustness intelligently. Extensive experiments on the COCO and SA-V benchmarks demonstrate that our method consistently outperforms state-of-the-art techniques.

In conclusion, this work establishes a principled pathway for the co-design of quantization and memory optimization for large vision models, showing the potential to improve their practicality significantly. We acknowledge current limitations, such as our importance metric being primarily optimized for linear layers. Future work could involve extending this co-design philosophy to more diverse architectures and a broader range of visual tasks.

Acknowledgments

This work is supported in part by the National Key Research and Development Program of China (No. 2023YFC3305600), Joint Fund of Ministry of Education of China (8091B022149, 8091B02072404), National Natural Science Foundation of China (62132016, 62571412, and 62571393), Key Research and Development Program of Shaanxi (2024GX-YBXM-127) and National Key Laboratory Foundation of China (Grant No. HTKJ2024KL504011).

References

Avinash Ramakanth, S.; and Venkatesh Babu, R. 2014. Seamseg: Video object segmentation using patch seams. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 376–383.

Barbiero, P.; Ciravegna, G.; Giannini, F.; Zarlenga, M. E.; Magister, L. C.; Tonda, A.; Lió, P.; Precioso, F.; Jamnik, M.; and Marra, G. 2023. Interpretable neural-symbolic concept reasoning. In *International Conference on Machine Learning*, 1801–1825. PMLR.

Choukroun, Y.; Kravchik, E.; Yang, F.; and Kisilev, P. 2019. Low-bit quantization of neural networks for efficient inference. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 3009–3018. IEEE.

Dettmers, T.; Pagnoni, A.; Holtzman, A.; and Zettlemoyer, L. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36: 10088–10115.

Ding, H.; Liu, C.; He, S.; Jiang, X.; Torr, P. H.; and Bai, S. 2023. MOSE: A new dataset for video object segmentation in complex scenes. In *Proceedings of the IEEE/CVF international conference on computer vision*, 20224–20234.

Dong, Z.; Yao, Z.; Gholami, A.; Mahoney, M. W.; and Keutzer, K. 2019. Hawq: Hessian aware quantization of neural networks with mixed-precision. In *Proceedings of the IEEE/CVF international conference on computer vision*, 293–302.

Frantar, E.; Ashkboos, S.; Hoefler, T.; and Alistarh, D. 2022. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*.

Hassibi, B.; Stork, D.; and Wolff, G. 1993. Optimal brain surgeon: Extensions and performance comparisons. *Advances in neural information processing systems*, 6.

Hu, Y.-T.; Huang, J.-B.; and Schwing, A. G. 2018. Videomatch: Matching based video object segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 54–70.

Huang, Q.; Guo, X.; Wang, Y.; Sun, H.; and Yang, L. 2024a. A survey of feature matching methods. *IET Image Processing*, 18(6): 1385–1410.

Huang, W.; Liu, Y.; Qin, H.; Li, Y.; Zhang, S.; Liu, X.; Magno, M.; and Qi, X. 2024b. Billm: Pushing the limit of post-training quantization for llms. *arXiv preprint arXiv:2402.04291*.

Jacob, B.; Kligys, S.; Chen, B.; Zhu, M.; Tang, M.; Howard, A.; Adam, H.; and Kalenichenko, D. 2018. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2704–2713.

Khan, S.; Naseer, M.; Hayat, M.; Zamir, S. W.; Khan, F. S.; and Shah, M. 2022. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s): 1–41.

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.

Lee, J.; Kim, D.; and Ham, B. 2021. Network quantization with element-wise gradient scaling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6448–6457.

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.; Xu, S.; Zhang, B.; Cao, X.; Gao, P.; and Guo, G. 2022. Q-vit: Accurate and fully quantized low-bit vision transformer. *Advances in neural information processing systems*, 35: 34451–34463.

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

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 *European conference on computer vision*, 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.

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.

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.

Nagel, M.; Fournarakis, M.; Amjad, R. A.; Bondarenko, Y.; Van Baalen, M.; and Blankevoort, T. 2021. A white paper on neural network quantization. *arXiv preprint arXiv:2106.08295*.

Perazzi, F.; Pont-Tuset, J.; McWilliams, B.; Van Gool, L.; Gross, M.; and Sorkine-Hornung, A. 2016. A benchmark dataset and evaluation methodology for video object segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 724–732.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rajbhandari, S.; Rasley, J.; Ruwase, O.; and He, Y. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, 1–16. IEEE.

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, X.; Li, X.; Wei, K.; Yang, X.; and Yang, Y. 2025. Q-MiniSAM2: A Quantization-based Benchmark for Resource-Efficient Video Segmentation. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, 1829–1837.

Shu, H.; Li, W.; Tang, Y.; Zhang, Y.; Chen, Y.; Li, H.; Wang, Y.; and Chen, X. 2025. Tinysam: Pushing the envelope for efficient segment anything model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 20470–20478.

Sun, M.; Liu, Z.; Bair, A.; and Kolter, J. Z. 2023. A simple and effective pruning approach for large language models. *arXiv preprint arXiv:2306.11695*.

Tang, C.; Ouyang, K.; Wang, Z.; Zhu, Y.; Ji, W.; Wang, Y.; and Zhu, W. 2022. Mixed-precision neural network quantization via learned layer-wise importance. In *European conference on computer vision*, 259–275. Springer.

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*.

Wu, H.; Judd, P.; Zhang, X.; Isaev, M.; and Micikevicius, P. 2020. Integer quantization for deep learning inference: Principles and empirical evaluation. *arXiv preprint arXiv:2004.09602*.

Wu, K.; Zhang, J.; Peng, H.; Liu, M.; Xiao, B.; Fu, J.; and Yuan, L. 2022. Tinyvit: Fast pretraining distillation for small vision transformers. In *European conference on computer vision*, 68–85. Springer.

Yang, Z.; Wei, Y.; and Yang, Y. 2021. Associating objects with transformers for video object segmentation. *Advances in Neural Information Processing Systems*, 34: 2491–2502.

Zhang, B.; Li, K.; Cheng, Z.; Hu, Z.; Yuan, Y.; Chen, G.; Leng, S.; Jiang, Y.; Zhang, H.; Li, X.; et al. 2025. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*.

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*.

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*.

Zhao, Y.; Lyu, G.; Li, K.; Wang, Z.; Chen, H.; Yang, Z.; and Deng, Y. 2025. ESEG: Event-Based Segmentation Boosted by Explicit Edge-Semantic Guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 10510–10518.

Zhu, K.; He, Y.-Y.; and Wu, J. 2023. Quantized feature distillation for network quantization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 11452–11460.