Q-MiniSAM2: A Quantization-based Benchmark for Resource-Efficient Video Segmentation

Xuanxuan Ren , Xiangyu Li , Kun Wei , Xu Yang* , Yanhua Yang*
Xidian University
{xxren, xdu_xyLi}@stu.xidian.edu.cn, {xuyang.xd, weikunsk}@gmail.com, yanhyang@xidian.edu.cn

Abstract

Segment Anything Model 2 (SAM2) is a new-generation, high-precision model for image and video segmentation, offering extensive application prospects across numerous computer vision fields. However, as a large-scale model, its huge memory demands and expansive computing costs pose challenges for practical deployment. This paper presents Q-MiniSAM2, an efficient ***Q***uantization-based segmentation benchmark tailored to optimize ***SAM2*** by ***Mini***mizing memory consumption and accelerating computations. We begin with applying Post-Training Quantization (PTQ) to SAM2, requiring only a relatively small dataset for network calibration, thereby eliminating the need for retraining. Building upon PTQ, we further introduce a Hierarchy-based Video Quantization method to enhance the model's capacity to capture video semantics and temporal correlations across different time scales. Furthermore, we observe that SAM2's memory overhead is predominantly concentrated on processing historical frames, and the redundant cross-attention computations significantly increase memory and computational costs due to the imperceptible change of the short time intervals between these frames. To tackle this issue, an Adaptive Mutual-KV mechanism is proposed to mitigate excessive cross-attention by leveraging inter-frame similarities. Comprehensive experiments demonstrate that the proposed approach achieves superior performance compared to state-of-the-art methods, underscoring its potential for efficient and scalable video segmentation.

1 Introduction

Large pre-trained models [Bi *et al.*, 2024; Xu *et al.*, 2024a] are capable of learning complex patterns from vast datasets, driving significant advances in natural language processing, computer vision, and computational biology. Their ability to generalize across domains enables adaptation to various downstream tasks directly or with minimal fine-tuning, without extensive retraining. As a prominent example, the Segment Anything Model (SAM) [Kirillov *et al.*, 2023], leverages large amounts of labeled and unlabeled data to learn generalized features, achieving high-precision segmentation across diverse image types, including previously unseen objects. Building on its predecessor, SAM2 [Ravi *et al.*, 2024] enhances image and video segmentation capabilities by integrating advanced techniques in adaptive prompt handling, multi-stage distillation, and optimized model architecture. Despite SAM's success, deploying it on the edge at scale presents challenges, such as managing computational costs and ensuring robustness across varied environments.

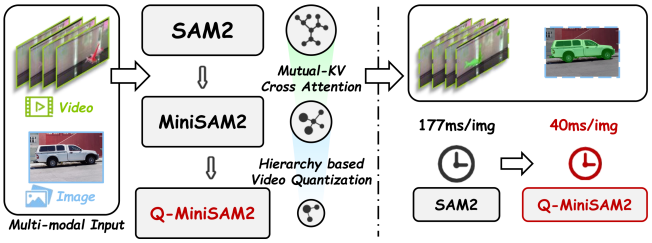

Figure 1: The existing edge quantization model can only perform image segmentation tasks. We proposed a new benchmark for the quantization segmentation model, which can be used for edge image and video segmentation tasks. At the same time, it has a lightweight design to address the timing challenges of videos.

Recent efforts have concentrated on advancing the deployment and application of SAM on edge devices. MobileSAM [Zhang *et al.*, 2023a] aims to reduce the computational load of the image encoder by adopting a lightweight TinyViT [Wu *et al.*, 2022] architecture, whereas FastSAM [Zhao *et al.*, 2023] reframes the segmentation task as an instance segmentation task with a single foreground category, utilizing the YOLOv8 [Jocher *et al.*, 2023] model. TinySAM [Shu *et al.*, 2023] introduces a comprehensive knowledge distillation approach that employs hard prompt sampling and a hard mask weighting strategy to train a lightweight student model. In PTQ4SAM [Lv *et al.*, 2024], Post-Training Quantization (PTQ) [Nagel *et al.*, 2020] is introduced to quantize SAM. Quantization aims to convert the floating-point parameters of the model into low-bit representation, which is a quantization method that requires only a small unlabeled dataset to cali-

*Corresponding author.

brate the network without retraining [Xu *et al.*, 2024b]. Unfortunately, all these methodologies are circumscribed within the ambit of the image domain, with scant incursion into the far more intricate expanse of the video domain.

As an innovative augmentation of SAM within the video realm, SAM2 integrates intricate real-time encoding and mask-generation faculties. Endowed with such characteristics, SAM2 manifests extraordinary suitability in applications of video and streaming data. It is not only capable of exquisitely segmenting objects but also exhibits remarkable resilience in accommodating diverse input modalities. Nonetheless, the prodigious quantity of parameters and the innate intricacy of video data present arduous challenges to the model's efficient inferential processes and edge-based deployment endeavors.

In the pursuit of mitigating the computational resource demands of SAM2 during edge deployment and video inference, we have devised an innovative approach: Q-MiniSAM2, a highly efficient quantization-based segmentation benchmark scheme, as graphically illustrated in Figure 1. This novel solution represents a significant step forward in optimizing the resource-intensive operations associated with SAM2. Specifically, we have been the first to proactively apply Post-Training Quantization (PTQ) technology to SAM2. In parallel, we have harnessed the power of hierarchy-based contrastive learning methods [Guo *et al.*, 2024; Li *et al.*, 2022a; Zhao *et al.*, 2021] to extract temporal information from relatively recent historical frames and semantic information from relatively distant historical frames. This dual-extraction mechanism ensures that the quantized model is exquisitely attuned to both the temporal coherence and semantic consistency between frames, which is crucial for accurate video segmentation.

Moreover, an in-depth analysis of video data reveals its pronounced redundancy in the temporal dimension. The changes between adjacent frames are trivial, leading to the unnecessary occupation of memory and the profligate waste of computational resources. To counteract this issue, we introduce the Adaptive Mutual-KV mechanism. During cross-attention, this mechanism enables the sharing of key-value (KV) pairs between similar historical frames and the current frame. This innovative approach effectively mitigates the high redundancy of video data in the temporal dimension, thereby enhancing the overall efficiency of the segmentation process. Our comprehensive approach has yielded remarkable results. It has successfully reduced the number of parameters of the original model by a staggering 80%, all while maintaining an impressive 93% of the original model's performance. This achievement represents a significant leap in the field of video segmentation, offering a more resource-efficient and cost-effective solution. In summary, our main contributions are as follows:

1. To the best of our knowledge, our work is the first to present a model quantization solution meticulously tailored for video segmentation models. By applying an adaptive quantization technology to SAM2, we have not only streamlined the computational process but also opened up new possibilities for optimizing video-based deep-learning models. This approach sets a new precedent in the domain, offering a novel paradigm for future research in resource-efficient video segmentation.
2. We introduce an innovative Hierarchy-based Video Quantization method to capture both the temporal coherence and semantic diversity from historical frames by leveraging contrastive learning. To address the redundancy in the representation of historical frames within the video segmentation model, we employ the adaptive mutual-KV mechanism, which can significantly reduce computational costs while maintaining the integrity of the segmentation process.
3. Extensive experimental results demonstrate that our proposed method consistently outperforms existing state-of-the-art PTQ approaches. Remarkably, our method maintains 93% of the original model's performance when compressed to 6-bit precision, and achieves state-of-the-art results even under aggressive 4-bit quantization.

2 Related Work

2.1 Video Object Segmentation

Video Object Segmentation (VOS) [Ding *et al.*, 2023; Yang *et al.*, 2021; Zhao *et al.*, 2025] is intended to segment a target foreground object from the video background at the pixel level across a video sequence. In Semi-Supervised VOS [Wang *et al.*, 2018; Wei *et al.*, 2019; Xu *et al.*, 2025], the process starts with an initial object mask in the first frame and requires accurate tracking and segmentation of the object throughout the video. Early neural network-based methods [Perazzi *et al.*, 2016; Xu *et al.*, 2018] typically employ online fine-tuning on the first frame of a video to adapt the model to the target object. Recent approaches [Wu *et al.*, 2023; Zhang *et al.*, 2023b] have extended single vision transformers to jointly process the current frame along with all previous frames and their associated predictions, achieving a simpler architecture but at the cost of significantly increased inference overhead.

For Interactive Video Object Segmentation (iVOS), the iVOS model segments objects (referred to as masklets) in the video based on user guidance, such as clicks or bounding boxes. Several early approaches [Wang *et al.*, 2005; Bai and Sapiro, 2007; Tao *et al.*, 2022] employ graph-based optimization techniques to guide the segmentation annotation process. More recent methods [Heo *et al.*, 2020; Cheng *et al.*, 2021; Delatolas *et al.*, 2024] often adopt a modular design, converting user input into a mask representation on a single frame and then propagating it to other frames. In SAM2, Semi-Supervised VOS and iVOS are unified into Promptable Visual Segmentation (PVS), enabling interactive segmentation and tracking of objects using inputs such as clicks, boxes, or masks on any video frame.

2.2 Model Quantization

Model quantization [Jacob *et al.*, 2018; Nagel *et al.*, 2021; Lyu *et al.*, 2020] reduces the numerical precision of model parameters, typically by converting floating-point values to

fixed-point ones. The aim is to reduce memory usage, computational complexity, and energy consumption while maintaining acceptable accuracy [Bi *et al.*, 2025]. Current research is predominantly categorized into Quantization-Aware Training (QAT) and Post-Training Quantization (PTQ).

Quantization-Aware Training (QAT) [Lee *et al.*, 2021; Li *et al.*, 2022b; Zhu *et al.*, 2023] integrates quantization into the training process through back-propagation, where a straight-through estimator [Bengio *et al.*, 2013] is commonly used to approximate the gradients of non-differentiable rounding functions. QAT involves retraining the model using the entire labeled dataset to optimize quantization parameters. However, given the massive scale of the original dataset, this approach would be highly time-consuming and computationally expensive. In contrast, PTQ [Frantar *et al.*, 2022; Li *et al.*, 2023; Dettmers *et al.*, 2024; Huang *et al.*, 2024] presents a more efficient alternative, as it only requires a small unlabeled dataset to calibrate the network. Adaround [Nagel *et al.*, 2020] proposes an adaptive weight rounding mechanism. Brecq [Li *et al.*, 2021] quantizes the model block by block and introduces the Fisher Information Matrix to guide the reconstruction process. Qdrop [Wei *et al.*, 2022] randomly discards activation quantization during the quantization process to improve the robustness of the quantization model.

3 Method

We first introduce the workflow of SAM2. In SAM2, videos are processed in a streaming manner, where each frame is sequentially processed by the image encoder and cross-attended with memory representations of the target object from prior frames by the memory attention. The mask decoder can optionally incorporate prompt inputs to predict the segmentation mask for the current frame. Lastly, the memory encoder transforms both the predictions and embeddings from the image encoder, preparing them for use in subsequent frames. In the following sections, we will elaborate on our method.

3.1 Video Segmentation Quantization Benchmark

Model quantization has become a crucial technique for accelerating and compressing deep learning models, as it allows for a reduction in the bit-level representation of parameters while maintaining model accuracy. In the most common quantization methods, quantization and de-quantization operations can be defined as follows:

$$\begin{aligned} \text{int} =& \texttt{clamp}\left(\left\lfloor \frac{\text{fp}}{s} \right\rceil + z, 0, 2^k - 1\right), \\ \hat{\text{fp}} =& s \cdot (\text{int} - z) \approx \text{fp}, \end{aligned} \tag{1}$$

where s and z denote the scaling factor and zero point, respectively. $\lfloor \cdot \rceil$ is the round-to-nearest operator. fp and $\hat{\text{fp}}$ are floating-point and de-quantized values, and int is mapped integer. `clamp` function clips the values fall outside the range of a k-bit integer.

In our analysis of the SAM2, we observed that the parameters of the image encoder, memory attention, and mask decoder modules collectively account for over 97% of the total model parameters. Given the substantial footprint of these three modules, we focus our quantization efforts on them to achieve the greatest impact in terms of memory and computational savings. To this end, we employ Post-Training Quantization, a strategy that requires significantly less computational effort compared to Quantization-Aware Training, as it bypasses the need for end-to-end model retraining.

Algorithm 1 Post-Training Quantization

Input: Full-precision SAM2 M_F, Calibration Set $Calib$
Output: Quantized model M_Q
Variables: $Qlayer$: layer in M_Q, $layer$: layer in M_F,s: Scale factor for quantization, z: Zero-point for quantization

1: Copy and insert fake quantization factors into M_F to obtain M_Q
2: Enable the observer for the fake quantization factors
3: $Calib$ is passed to M_Q for calibration, and the observer collects activation distribution
4: Enable the fake-quant for the fake quantization factors
5: **for** $Qlayer$ in M_Q, $layer$ in M_F **do**
6: Update the parameters s and z by Eq. 2 and Eq. 3
7: **end for**
8: **return** M_Q

Our post-training quantization method introduces fake quantization factors into the full-precision SAM2 model, denoted as M_F, to obtain the quantized model M_Q. We then enable the observer for the fake quantization factors, and a small calibration dataset, $Calib$, is used to evaluate and record the activation distributions within each layer under simulated quantization conditions. This calibration phase is crucial, as it allows precise capture of each layer's output statistics, such as activation range and distribution, which are used to initialize the quantization scaling factors. The calibration data, $Calib$, then guides a layer-by-layer reconstruction of the quantized model, aiming to closely align the quantized outputs with those of the original floating-point model. During reconstruction, loss functions are applied to minimize the gap between the outputs of the quantized and original models. Similar to AdaRound [Nagel *et al.*, 2020], We define the reconstruction loss as:

$$\mathcal{L}_{rec} = \left\| \mathbf{W}\mathbf{x} - \widetilde{\mathbf{W}}\mathbf{x} \right\|_F^2 + \alpha R(\mathbf{V}), \tag{2}$$

where the first term represents the MSE loss, $\mathbf{W}$ is the weights of the reconstruction layer. $\mathbf{x}$ is the input of the reconstruction layer, α is a trade-off parameter and $\widetilde{\mathbf{W}}$ are the quantized weights:

$$\widetilde{\mathbf{W}} = \texttt{clamp}(\lfloor \frac{\mathbf{W}}{s} \rfloor + h(\mathbf{V}) + z, 0, 2^k - 1). \tag{3}$$

The other term $R(\mathbf{V})$ is a differentiable regularizer that is encouraged the optimization variables $h(\mathbf{V}_{i,j})$ to converge towards either 0 or 1. The overall process is shown in Algorithm 1.

3.2 Hierarchy-based Video Quantization

In our Video Segmentation Quantization Benchmark, to minimize the accuracy loss between the quantized model and the

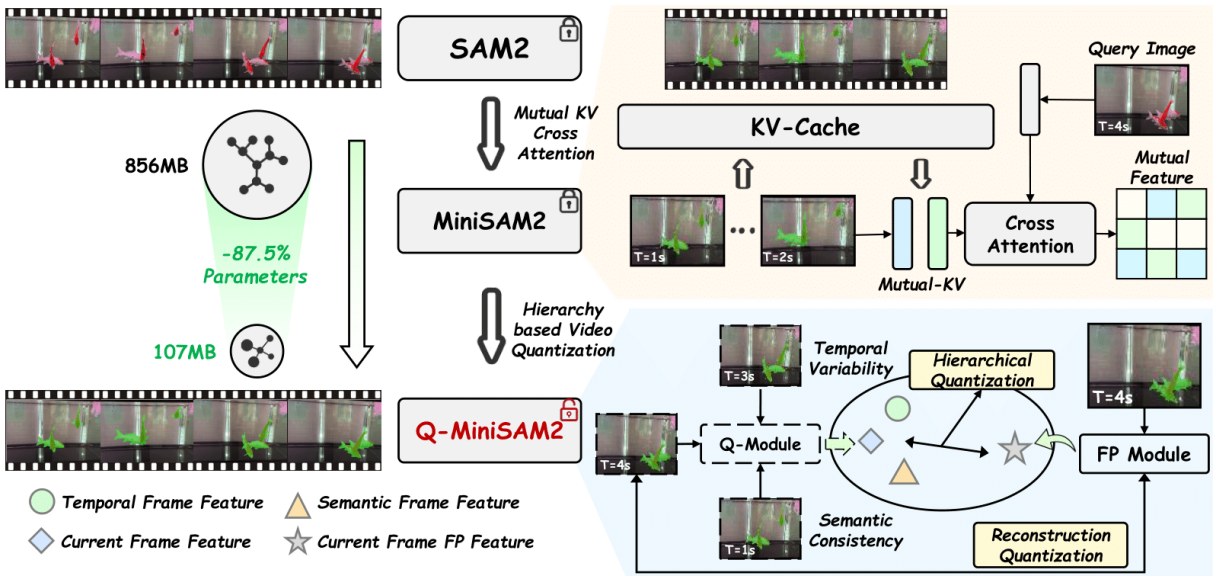

Figure 2: Illustration of our proposed Q-MiniSAM2. The Adaptive Mutual-KV selectively shares the computation of Key (K) and Value (V) matrix representations to reduce computational costs without sacrificing accuracy. The Hierarchy-based Video Quantization (HVQ) constructs multi-scale representations by leveraging temporal variability and semantic consistency, ensuring high-quality quantization through multi-scale reconstruction.

floating-point model, we utilize Eq. 2 to optimize the quantization parameters, ensuring that the outputs of the floating-point model and the quantized model for the same frame are closely aligned. However, Eq. 2 only considers output alignment. To fully account for the variations in both short-term and long-term frames in videos, we propose a novel Hierarchy-based Video Quantization framework. For the current frame, nearby historical frames are typically semantically similar and contain the temporal information we aim to capture; thus, we refer to these frames as $temporal\ frames$. On the other hand, more distant historical frames may experience significant semantic changes due to the longer time interval. In order to make more full use of the potential information of these frames, we hope to extract richer semantic features from them, which we call $semantic\ frames$.

Firstly, We use the output of the floating-point model for the current frame as an anchor point. We then minimize the distance between the anchor and the output of the quantized model for the current frame, while simultaneously maximizing the distances between the anchor and the outputs of the quantized models for both $temporal\ frames$ and $semantic\ frames$. This approach enables us to effectively capture the temporal information from nearby frames and the semantic information from distant frames.

Specifically, in each iteration, we select a frame t from the video as the current frame and choose a $temporal\ frame$ t_{tem} from the interval $(t - l_{\text{short}}, t)$ and a $semantic\ frame$ t_{sem} from the interval $(t - l_{\text{long}}, t - l_{\text{short}})$. l_{short} and l_{long} denote the lengths of the short-term and long-term time windows, respectively. The current frame feature $\mathbf{f}_t$ is then passed through both the reconstructed floating-point module and the quantized module to obtain their respective outputs denoted as $\mathbf{f}_t^{\text{fp\_out}}$ and $\mathbf{f}_t^{\text{q\_out}}$. Similarly, the $temporal\ frame$ feature $\mathbf{f}_{\text{tem}}$ and the $semantic\ frame$ feature $\mathbf{f}_{\text{sem}}$ are passed through the reconstructed quantized module to generate their outputs, $\mathbf{f}_{\text{tem}}^{\text{q\_out}}$ and $\mathbf{f}_{\text{sem}}^{\text{q\_out}}$. Using $\mathbf{f}_t^{\text{fp\_out}}$ as the anchor, we employ two triplet losses to minimize the distance between $\mathbf{f}_t^{\text{fp\_out}}$ and $\mathbf{f}_t^{\text{q\_out}}$, while simultaneously maximizing the distances between $\mathbf{f}_t^{\text{fp\_out}}$ and the outputs of the quantized module for the $temporal\ frame$ and $semantic\ frame$, $\mathbf{f}_{\text{tem}}^{\text{q\_out}}$ and $\mathbf{f}_{\text{sem}}^{\text{q\_out}}$. The formula is as follows:

$$\begin{aligned}\mathcal{L}_{\text{Hiera}} = \mathcal{L}_{\text{triplet}}\left(\mathbf{f}_t^{\text{fp\_out}}, \mathbf{f}_t^{\text{q\_out}}, \mathbf{f}_{\text{tem}}^{\text{q\_out}}\right) + \\ \beta \cdot \mathcal{L}_{\text{triplet}}\left(\mathbf{f}_t^{\text{fp\_out}}, \mathbf{f}_t^{\text{q\_out}}, \mathbf{f}_{\text{sem}}^{\text{q\_out}}\right),\end{aligned} \tag{4}$$

where β is a trade-off parameter. This approach ensures that the quantized model not only aligns closely with the floating-point model for the current frame but also captures the temporal consistency from nearby frames and the semantic diversity from distant frames.

Eventually, the total loss of our proposed framework is formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{Hiera}} + \gamma \cdot \mathcal{L}_{\text{rec}}, \tag{5}$$

where γ is a trade-off parameter.

3.3 Adaptive Mutual-KV

To address the computational complexity and memory consumption challenges in video segmentation models for video sequence processing, we propose a novel Adaptive Mutual-KV method based on the similarity of historical frames. In traditional Cross-Attention mechanisms, each historical frame independently generates a distinct Key-Value pair, leading to substantial increases in memory usage and computational overhead. Our method innovatively calculates the similarity between frames and shares the same Key-Value

Methods	Model-S			Model-B			Model-L		
	FP	W6A6	W4A4	FP	W6A6	W4A4	FP	W6A6	W4A4
MinMax [Jacob *et al.*, 2018]	40.3	10.9	-	41.1	35.5	-	41.4	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		**39.0**	**34.7**		**40.0**	**33.9**		**38.8**	**31.8**

Table 1: Results of image instance segmentation on COCO dataset among different methods. FP represents the original floating point model, W6A6 represents quantizing weights and activations to 6-bit, and W4A4 represents quantizing weights and activations to 4-bit. - indicates the final result (mAP) is below 1.

pair among similar historical frames, rather than generating a new Key-Value pair for each individual frame.

The core principle of this approach is to group historical frames based on their similarity, thereby enabling the sharing of Key-Value pairs within each group. This strategy not only reduces memory consumption but also ensures the retention of critical information across historical frames. By striking a balance between computational efficiency and information retention, our method becomes particularly advantageous for real-time video processing tasks, where both computational resources and memory are often constrained.

Our method begins by calculating the similarity between the historical frames. This similarity score is used to identify frames that are most similar to each other, and those with similarity scores above a dynamic threshold ϵ are grouped together. The formula is as follows:

$$G_t = \{i \mid \frac{\mathbf{f}_t \cdot \mathbf{f}_i}{\|\mathbf{f}_t\| \, \|\mathbf{f}_i\|} > \epsilon; i, t \in G\}, \tag{6}$$

where $\mathbf{f}_t$ and $\mathbf{f}_i$ are the feature of frames t and i, respectively. G is the set of all historical frames, G_t is the frame group obtained by frame t, ϵ is a dynamic threshold. In the early stages of training, a larger ϵ is used to prevent low-similarity frames from sharing the same Key-Value pair, ensuring information independence and avoiding loss. As training progresses and the model learns temporal dependencies, ϵ is gradually reduced to decrease computational cost.

Once the historical frames have been grouped based on similarity, we proceed to generate shared Key-Value pairs. For a group of similar frames, we calculate a collective $\mathbf{Key}_{G_t}$ and a collective $\mathbf{Value}_{G_t}$, which will be shared among all frames in the group G_t. Specifically, for a frames group G_t the shared $\mathbf{Key}_{G_t}$ and $\mathbf{Value}_{G_t}$ can be computed as the average of the individual Keys and Values:

$$\mathbf{Key}_{G_t} = \frac{1}{|G_t|} \sum_{j \in G_t} \mathbf{Key}_j, \mathbf{Value}_{G_t} = \frac{1}{|G_t|} \sum_{j \in G_t} \mathbf{Value}_j, \tag{7}$$

where $\mathbf{Key}_j$ and $\mathbf{Value}_j$ are the Key and Value for each frame j in the group G_t. Then, we use the obtained shared key-value in cross attention:

$$\text{Attention}(Q, K_j, V_j) = \text{softmax}\left(\frac{Q \cdot \mathbf{Key}_{G_t}^T}{\sqrt{d_k}}\right) \mathbf{Value}_{G_t}, \tag{8}$$

where Q is the query obtained from the current frame, K_j, V_j are the key and value obtained from a certain historical frame j, and $j \in G_t$. This aggregation not only reduces the number of unique Key-Value pairs used in the Cross-Attention mechanism, but also ensures that the historical frame information is effectively preserved. By sharing the same Key-Value pair among similar frames, we retain the relevant contextual information of the historical frames while significantly improving computational efficiency.

4 Experiments

4.1 Experimental Setups

Datasets. We conduct experiments on two object segmentation datasets: MS-COCO [Lin *et al.*, 2014] and SA-V [Ravi *et al.*, 2024]. MS-COCO contains 123,000 images across 91 object categories, of which the training set contains 118,000 images and the validation set containing 5,000 images. The SA-V dataset comprises approximately 51,000 real-world videos and over 600,000 spatiotemporal masks (referred to as masklets), establishing it as the largest video segmentation dataset to date. Specifically, the training split consists of 505,83 videos and 642,036 masklets, while the validation split includes 155 videos and 293 masklets. Additionally, the test split contains 150 videos and 278 masklets.

Tasks and metrics. We conduct experiments on two segmentation tasks. For Promptable Visual Segmentation (PVS), we obtain accurate object masks by manually annotating box prompts and use $\mathcal{J}\&\mathcal{F}$ (Jaccard and F-measure) to evaluate its effectiveness on the SA-V dataset. $\mathcal{J}\&\mathcal{F}$ combines the Jaccard Index ($\mathcal{J}$) and F-measure ($\mathcal{F}$) to evaluate the overlap area and boundary accuracy of the segmentation results, respectively. For the image instance segmentation task, we leverage the predicted bounding boxes generated by the detector as box prompts for the SAM2 to obtain precise binary masks. To evaluate the effectiveness of this approach, we

Methods	Model-B									Model-L								
	FP			W6A6			W4A4			FP			W6A6			W4A4		
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Adaround [Nagel *et al.*, 2020]	72.7	69.0	76.3	47.2	43.5	50.9	25.6	23.8	27.5	73.7	70.3	77.2	48.1	44.5	51.7	27.3	25.7	28.9
Qdrop [Wei *et al.*, 2022]				61.4	57.7	65.0	33.4	30.7	36.1				63.1	59.3	66.8	35.8	33.7	37.9
PTQ4SAM [Lv *et al.*, 2024]				66.5	63.0	69.9	37.8	35.6	40.0				67.1	63.6	70.5	39.5	37.4	41.6
Ours				**67.9**	**64.2**	**71.5**	**39.1**	**36.5**	**41.6**				**68.8**	**65.1**	**72.4**	**40.7**	**38.5**	**42.9**

Table 2: Results of promptable visual segmentation on SA-V dataset among different methods.

Methods	Model-B		
	FP	W6A6	W4A4
Base	72.7	66.6	38.0
Base+MKV		67.3	38.5
Base+HVQ		67.1	38.7
Ours		**67.9**	**39.1**

Table 3: Ablation Sdudies of promptable visual segmentation on SA-V dataset ($\mathcal{J}\&\mathcal{F}$).

employ the mean Average Precision (mAP) as the primary evaluation metric.

Implementation details. In the image instance segmentation task, YOLOX [Ge, 2021] is employed as the detector to generate predicted bounding boxes, which serve as the box prompt inputs for the SAM2 model. For quantization training, a set of 32 unannotated training images is randomly selected to form the training dataset. In the prompt-based visual segmentation task, to obtain accurate target masks through manually annotated box prompts, 8 videos are randomly chosen from the SA-V validation set, with 20 frames extracted from each video to construct the training dataset. Following conventional methodologies, the implemented quantization strategy includes per-channel asymmetric quantization for weights and per-tensor asymmetric quantization for activation values. Each module undergoes 20,000 iterations during the reconstruction phase. Additionally, to ensure the stability and robustness of the model's performance, the first and last layers (or modules) of the network are exempted from the quantization process. The hyperparameters α, β, and γ are set to 1, 0.5 and 0.4 respectively.

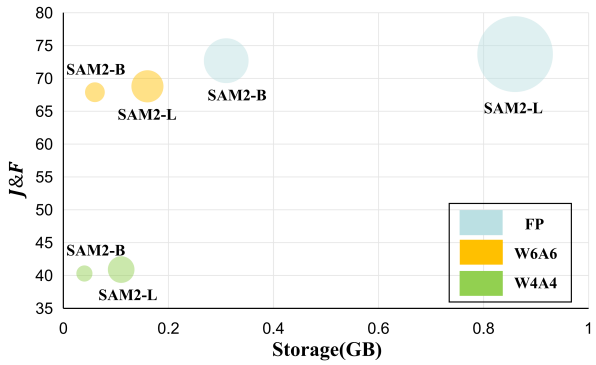

Figure 3: SAM2 performance at different scales and compression levels

4.2 Image Instance Segmentation Results

We compare Q-MiniSAM2 with statistical-based methods (e.g., MinMax [Jacob *et al.*, 2018], Percentile [Wu *et al.*, 2020], OMSE [Choukroun *et al.*, 2019]) and learning-based methods (e.g., AdaRound [Nagel *et al.*, 2020], BRECQ [Li *et al.*, 2021], QDrop [Wei *et al.*, 2022]) under the same setting. As shown in Table 1, our method consistently achieves superior performance compared to the other approaches, demonstrating a significant improvement in evaluated metrics. Our 4-bit Q-MiniSAM2 is 16.3% mAP higher than the current best method PTQ4SAM on SAM-S and achieves the best results on SAM-B and SAM-L. As the model size increases from SAM2-B to SAM2-L, the segmentation performance of the quantized model degrades. We attribute this phenomenon to the fact that the error accumulation caused by the increased number of layers outweighs the performance gains from the larger model capacity. Notably, even for the full-precision (FP) model, the performance improvement from SAM2-B to SAM2-L is marginal, with only a 0.3% increase in mAP. Specifically, as the depth of the model increases, quantization errors will gradually propagate and accumulate in the network. This phenomenon will appear when the performance improvement of the model cannot make up for the error introduced by quantization.

4.3 Promptable Visual Segmentation Results

We apply AdaRound, QDrop and PTQ4SAM to SAM2 and compare their performance with our proposed method, Q-MiniSAM2. The experimental results are summarized in Table 2. Our method maintains 93% of the original model's performance even when compressed to 6-bit precision. For instance, on SAM2-B, the performance drop is only 4.8% in terms of $\mathcal{J}\&\mathcal{F}$. Compared to the state-of-the-art image quantization segmentation method, PTQ4SAM, our method achieves a significant improvement, outperforming it by 1.4% in $\mathcal{J}\&\mathcal{F}$ at 6-bit precision. In the case of 4-bit quantization for SAM2-B, where the parameter precision is significantly reduced, our method still achieves 39.1% in $\mathcal{J}\&\mathcal{F}$, making it the best-performing video segmentation quantization method to date.

4.4 Ablation Studies

We conduct an ablation study to evaluate the effectiveness of the proposed Adaptive Mutual-KV (MKV) and Hierarchy-based Video Quantization (HVQ). Using the quantized SAM2 framework described in Section 3.2 as the baseline (Base), we introduce two variants by integrating Adaptive Mutual-KV and Hierarchy-based Video Quantization separately. Experiments are conducted on the SA-V dataset,

Figure 4: Qualitative results of promptable visual segmentation on the SA-V dataset.

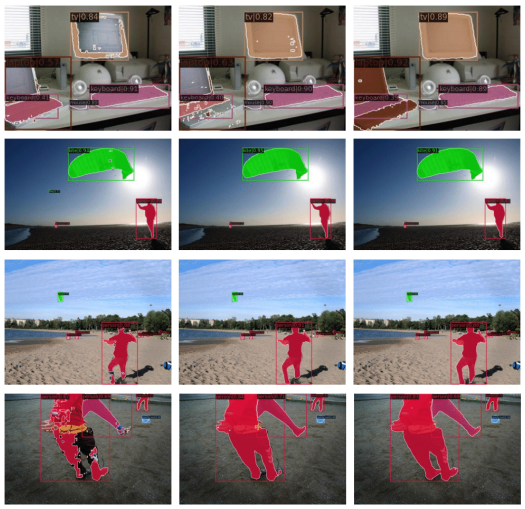

Figure 5: Qualitative results of image instance segmentation

and the results are summarized in Table 3. As shown in Table 3, each variant outperforms the baseline, demonstrating the individual contributions of Adaptive Mutual-KV and Hierarchy-based Video Quantization. Furthermore, the combination of both components achieves the best performance, indicating that these components complement each other and work synergistically to significantly enhance performance. Specifically, our proposed the Hierarchy based Video Quantization successfully captures both temporal variability and semantic consistency in video data, enabling high-quality quantization. The Adaptive Mutual-KV enhances model performance while reducing computational costs.

4.5 Storage Saving

We evaluated the performance of SAM2 models of varying sizes when compressed to different bit precisions using Q-MiniSAM2, as illustrated in Figure 3. At W6A6 precision, our method reduces storage by over 80%, while retaining 93% of the original performance. On specialized hardware, with W4A4 precision, the inference speed can theoretically increase by a factor of 8 compared to the original model. This is because the computation of a 32-bit multiplication can now be replaced by performing 8 parallel 4-bit multiplications. While factors such as quantization parameters may influence the actual results, the expected inference speed is at least 5 times faster than that of the full-precision model.

4.6 Qualitative Results

In Figure 4, we present the qualitative results of promptable visual segmentation. From the figure, it is evident that our method not only generates fine-grained object masks from the video (e.g., the body and tail of the bird in the first row), but also maintains continuous tracking of occluded objects (e.g., the person occluded in the third row). This is made possible by the improvements we have specifically designed for video processing. In Figure 5, we present the qualitative results of image instance segmentation. Compared to existing methods, it is evident that most approaches fail to produce clear object boundaries (as seen in the ”person” in the fourth row) and suffer mask omissions (such as the ”monitor” in the first row). Furthermore, many methods struggle to segment objects in complex scenes (e.g., the ”laptop” in the first row). In contrast, our method clearly outperforms others, demonstrating superior mask completeness and boundary clarity.

5 Conclusion

In this paper, we propose a novel post-training quantization framework, Q-MiniSAM2, for video segmentation. First, we establish a post-training quantization benchmark for SAM2. Second, we observe redundancy in the interaction between the current frame and historical frames in SAM2, and we introduce Adaptive Mutual-KV to reduce redundancy across historical frames. Then, to capture the temporal variability and semantic consistency in videos, we propose Hierarchy-based Video Quantization, which simultaneously considers the temporal relationships of nearby frames and the semantic information of distant frames to ensure high-quality quantization. Extensive experimental results demonstrate the effectiveness and practicality of our method. However, we note that performance degradation at 4-bit quantization remains significant, and this will be a focus of our future work.

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 (8091B02072404), National Natural Science Foundation of China (62132016, 62171343, and 62201436), Key Research and Development Program of Shaanxi (2024GX-YBXM-127), Natural Science Basic Research Program of Shaanxi (2020JC-23) and National Key Laboratory Foundation of China (Grant No. HTKJ2024KL504011).

Contribution Statement

Xuanxuan Ren and Xiangyu Li contribute equally. They jointly designed the experiments and wrote the manuscript. Xu Yang and Kun Wei contributed to the writing of the manuscript. Yanhua Yang and Xu Yang supervise the project.

References

[Bai and Sapiro, 2007] Xue Bai and Guillermo Sapiro. A geodesic framework for fast interactive image and video segmentation and matting. In *ICCV*, pages 1–8. IEEE, 2007.

[Bengio *et al.*, 2013] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

[Bi *et al.*, 2024] Jinhe Bi, Yujun Wang, Haokun Chen, Xun Xiao, Artur Hecker, Volker Tresp, and Yunpu Ma. Visual instruction tuning with 500x fewer parameters through modality linear representation-steering. *arXiv preprint arXiv:2412.12359*, 2024.

[Bi *et al.*, 2025] Jinhe Bi, Yifan Wang, Danqi Yan, Xun Xiao, Artur Hecker, Volker Tresp, and Yunpu Ma. Prism: Self-pruning intrinsic selection method for training-free multimodal data selection. *arXiv preprint arXiv:2502.12119*, 2025.

[Cheng *et al.*, 2021] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Modular interactive video object segmentation: Interaction-to-mask, propagation and difference-aware fusion. In *CVPR*, pages 5559–5568, 2021.

[Choukroun *et al.*, 2019] Yoni Choukroun, Eli Kravchik, Fan Yang, and Pavel Kisilev. Low-bit quantization of neural networks for efficient inference. In *ICCVW*, pages 3009–3018. IEEE, 2019.

[Delatolas *et al.*, 2024] Thanos Delatolas, Vicky Kalogeiton, and Dim P Papadopoulos. Learning the what and how of annotation in video object segmentation. In *WACV*, pages 6951–6961, 2024.

[Dettmers *et al.*, 2024] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *NeurIPS*, 36, 2024.

[Ding *et al.*, 2023] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Philip HS Torr, and Song Bai. Mose: A new dataset for video object segmentation in complex scenes. In *ICCV*, pages 20224–20234, 2023.

[Frantar *et al.*, 2022] Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*, 2022.

[Ge, 2021] Z Ge. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

[Guo *et al.*, 2024] Ruiming Guo, Mouxing Yang, Yijie Lin, Xi Peng, and Peng Hu. Robust contrastive multi-view clustering against dual noisy correspondence. *NeurIPS*, 37:121401–121421, 2024.

[Heo *et al.*, 2020] Yuk Heo, Yeong Jun Koh, and Chang-Su Kim. Interactive video object segmentation using global and local transfer modules. In *ECCV*, pages 297–313. Springer, 2020.

[Huang *et al.*, 2024] Wei Huang, Yangdong Liu, Haotong Qin, Ying Li, Shiming Zhang, Xianglong Liu, Michele Magno, and Xiaojuan Qi. Billm: Pushing the limit of post-training quantization for llms. *arXiv preprint arXiv:2402.04291*, 2024.

[Jacob *et al.*, 2018] 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 *CVPR*, pages 2704–2713, 2018.

[Jocher *et al.*, 2023] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Yolo by ultralytics. 2023.

[Kirillov *et al.*, 2023] 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 *ICCV*, pages 4015–4026, 2023.

[Lee *et al.*, 2021] Junghyup Lee, Dohyung Kim, and Bumsub Ham. Network quantization with element-wise gradient scaling. In *CVPR*, pages 6448–6457, 2021.

[Li *et al.*, 2021] Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Hu, Qi Zhang, Fengwei Yu, Wei Wang, and Shi Gu. Brecq: Pushing the limit of post-training quantization by block reconstruction. *arXiv preprint arXiv:2102.05426*, 2021.

[Li *et al.*, 2022a] Xiangyu Li, Xu Yang, Kun Wei, Cheng Deng, and Muli Yang. Siamese contrastive embedding network for compositional zero-shot learning. In *CVPR*, pages 9326–9335, 2022.

[Li *et al.*, 2022b] Yanjing Li, Sheng Xu, Baochang Zhang, Xianbin Cao, Peng Gao, and Guodong Guo. Q-vit: Accurate and fully quantized low-bit vision transformer. *NeurIPS*, 35:34451–34463, 2022.

[Li *et al.*, 2023] Zhikai Li, Junrui Xiao, Lianwei Yang, and Qingyi Gu. Repq-vit: Scale reparameterization for post-training quantization of vision transformers. In *ICCV*, pages 17227–17236, 2023.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco:

Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.

[Lv *et al.*, 2024] Chengtao Lv, Hong Chen, Jinyang Guo, Yifu Ding, and Xianglong Liu. Ptq4sam: Post-training quantization for segment anything. In *CVPR*, pages 15941–15951, 2024.

[Lyu *et al.*, 2020] Gengyu Lyu, Songhe Feng, Yidong Li, Yi Jin, Guojun Dai, and Congyan Lang. Hera: partial label learning by combining heterogeneous loss with sparse and low-rank regularization. *TIST*, 11(3):1–19, 2020.

[Nagel *et al.*, 2020] Markus Nagel, Rana Ali Amjad, Mart Van Baalen, Christos Louizos, and Tijmen Blankevoort. Up or down? adaptive rounding for post-training quantization. In *ICML*, pages 7197–7206. PMLR, 2020.

[Nagel *et al.*, 2021] 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.

[Perazzi *et al.*, 2016] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, pages 724–732, 2016.

[Ravi *et al.*, 2024] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[Shu *et al.*, 2023] Han Shu, Wenshuo Li, Yehui Tang, Yiman Zhang, Yihao Chen, Houqiang Li, Yunhe Wang, and Xinghao Chen. Tinysam: Pushing the envelope for efficient segment anything model. *arXiv preprint arXiv:2312.13789*, 2023.

[Tao *et al.*, 2022] Renshuai Tao, Tianbo Wang, Ziyang Wu, Cong Liu, Aishan Liu, and Xianglong Liu. Few-shot x-ray prohibited item detection: A benchmark and weak-feature enhancement network. In *ACM MM*, pages 2012–2020, 2022.

[Wang *et al.*, 2005] Jue Wang, Pravin Bhat, R Alex Colburn, Maneesh Agrawala, and Michael F Cohen. Interactive video cutout. *ToG*, 24(3):585–594, 2005.

[Wang *et al.*, 2018] Wenguan Wang, Jianbing Shen, Fatih Porikli, and Ruigang Yang. Semi-supervised video object segmentation with super-trajectories. *TPAMI*, 41(4):985–998, 2018.

[Wei *et al.*, 2019] Kun Wei, Muli Yang, Hao Wang, Cheng Deng, and Xianglong Liu. Adversarial fine-grained composition learning for unseen attribute-object recognition. In *ICCV*, pages 3741–3749, 2019.

[Wei *et al.*, 2022] Xiuying Wei, Ruihao Gong, Yuhang Li, Xianglong Liu, and Fengwei Yu. Qdrop: Randomly dropping quantization for extremely low-bit post-training quantization. *arXiv preprint arXiv:2203.05740*, 2022.

[Wu *et al.*, 2020] Hao Wu, Patrick Judd, Xiaojie Zhang, Mikhail Isaev, and Paulius Micikevicius. Integer quantization for deep learning inference: Principles and empirical evaluation. *arXiv preprint arXiv:2004.09602*, 2020.

[Wu *et al.*, 2022] Kan Wu, Jinnian Zhang, Houwen Peng, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. Tinyvit: Fast pretraining distillation for small vision transformers. In *ECCV*, pages 68–85. Springer, 2022.

[Wu *et al.*, 2023] Qiangqiang Wu, Tianyu Yang, Wei Wu, and Antoni B Chan. Scalable video object segmentation with simplified framework. In *ICCV*, pages 13879–13889, 2023.

[Xu *et al.*, 2018] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *ECCV*, pages 585–601, 2018.

[Xu *et al.*, 2024a] Chenghao Xu, Guangtao Lyu, Jiexi Yan, Muli Yang, and Cheng Deng. Llm knows body language, too: Translating speech voices into human gestures. In *ACL*, pages 5004–5013, 2024.

[Xu *et al.*, 2024b] Chenghao Xu, Jiexi Yan, Muli Yang, and Cheng Deng. Rethinking noise sampling in class-imbalanced diffusion models. *TIP*, 2024.

[Xu *et al.*, 2025] Chenghao Xu, Jiexi Yan, and Cheng Deng. Keep and extent: Unified knowledge embedding for few-shot image generation. *TIP*, 2025.

[Yang *et al.*, 2021] Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation. *NeurIPS*, 34:2491–2502, 2021.

[Zhang *et al.*, 2023a] Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023.

[Zhang *et al.*, 2023b] Jiaming Zhang, Yutao Cui, Gangshan Wu, and Limin Wang. Joint modeling of feature, correspondence, and a compressed memory for video object segmentation. *arXiv preprint arXiv:2308.13505*, 2023.

[Zhao *et al.*, 2021] Han Zhao, Xu Yang, Zhenru Wang, Erkun Yang, and Cheng Deng. Graph debiased contrastive learning with joint representation clustering. In *IJCAI*, pages 3434–3440, 2021.

[Zhao *et al.*, 2023] Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, and Jinqiao Wang. Fast segment anything. *arXiv preprint arXiv:2306.12156*, 2023.

[Zhao *et al.*, 2025] Yucheng Zhao, Gengyu Lyu, Ke Li, Zihao Wang, Hao Chen, Zhen Yang, and Yongjian Deng. Eseg: Event-based segmentation boosted by explicit edge-semantic guidance. In *AAAI*, volume 39, pages 10510–10518, 2025.

[Zhu *et al.*, 2023] Ke Zhu, Yin-Yin He, and Jianxin Wu. Quantized feature distillation for network quantization. In *AAAI*, volume 37, pages 11452–11460, 2023.