# When W4A4 Breaks Camouflaged Object Detection: Token-Group Dual-Constraint Activation Quantization

Tianqi Li[1], Wenyu Fang[1], Xin He[2], Xue Geng[3], Xu Cheng[2], and Yun Liu[1,4,5]*

[1] VCIP, College of Computer Science, Nankai University
[2] School of Computer Science and Engineering, Tianjin University of Technology
[3] Institute for Infocomm Research, A*STAR
[4] Academy for Advanced Interdisciplinary Studies, Nankai University
[5] Nankai International Advanced Research Institute, Shenzhen Futian

**Abstract.** Camouflaged object detection (COD) segments objects that intentionally blend with the background, so predictions depend on subtle texture and boundary cues. COD is often needed under tight on-device memory and latency budgets, making low-bit inference highly desirable. However, COD is unusually hard to quantize aggressively. We study post-training W4A4 quantization of Transformer-based COD and find a task-specific cliff: heavy-tailed background tokens dominate a shared activation range, inflating the step size and pushing weak-but-structured boundary cues into the zero bin. This exposes a token-local bottleneck—remove cross-token *range domination* and bound the *zero-bin mass* under 4-bit activations. To address this, we introduce **COD-TDQ**, a **COD**-aware **T**oken-group **D**ual-constraint activation **Q**uantization method. COD-TDQ addresses this token-local bottleneck with two coupled steps: **D**irect-**S**um **T**oken-**G**roup (**DSTG**) assigns *token-group* scales to suppress cross-token range domination, and **D**ual-**C**onstraint **R**ange **P**rojection (**DCRP**) projects each token-group clip range to keep the step-to-dispersion ratio and the zero-bin mass bounded. Across four COD benchmarks and two baseline models (CFRN and ESCNet), COD-TDQ consistently achieves an $S_\alpha$ score more than 0.12 higher than that of the state-of-the-art quantization method without retraining. The code is available at https://github.com/MCG-NKU/nku-model-compre.

**Keywords:** Camouflaged Object Detection · Post-Training Quantization · Token-Group Quantization · Dual-Constraint Quantization

## 1 Introduction

Camouflaged object detection (COD) is typically evaluated as binary mask prediction for objects that intentionally blend into the background. COD models must rely on subtle texture and boundary cues, so useful evidence often appears

* Corresponding author: Yun Liu (liuyun@nankai.edu.cn)

as small-amplitude yet structured responses [43]. As Transformer [6, 29, 46] encoders and multi-stage designs become common [2,4,12,22,37,55,60], COD models have witnessed significant development [7, 15, 20, 23, 25, 35, 39, 40, 49, 59, 61]. However, with this growth, COD models are becoming more complex, increasing deployment memory and latency costs [5, 26, 44, 45, 53]. At the same time, COD is increasingly expected to operate under strict memory and latency constraints [10,13,41,48], especially for applications on mobile and edge devices. In this context, it becomes desirable to reduce the computational and storage costs of models during deployment. One practical solution is to apply post-training quantization (PTQ), which can effectively reduce memory usage and activation cost without requiring retraining [3].

In practice, INT8 PTQ is often the default choice [9,52]. Nevertheless, when deployment budgets are extremely limited, the gains from INT8 can be modest, and further reductions require moving to ultra-low-bit regimes. This motivates exploring ultra-low-bit PTQ for COD while maintaining accuracy. In particular, 4-bit weights and 4-bit activations (W4A4) reduce deployment cost under a standard protocol [3, 38] without changing the training or inference pipeline. However, we find that naive W4A4 on Transformer-based COD can collapse rather than degrade smoothly. This raises a key question: *what makes COD fragile at 4-bit activations, and how can we make W4A4 reliable without retraining or hardware-specific assumptions?*

Transformer PTQ has advanced rapidly, but most pipelines still fall into a few recurring designs: (i) block-wise reconstruction and rounding to match FP32 outputs on calibration data [24, 50, 56], (ii) range re-parameterization and smoothing to mitigate heavy-tailed activations [51], and (iii) finer granularity via grouping or adaptive scales [33]. These strategies are largely layer or block-centric and optimize average fidelity. In COD at W4A4, the dominant failure is token-local: token-wise activation heterogeneity lets background tokens dominate the range and increase the zero-bin mass, which layer-wise fitting or reconstruction does not explicitly bound.

We trace the W4A4 cliff to a coupled collapse mechanism of **range domination** and **zero-bin mass** (Fig. 1). Background tokens with heavy-tailed activation spikes can dominate a shared clipping range, inflating the step size $\Delta$ and coarsening quantization for most tokens (Fig. 1). Under rounding-to-nearest, weak yet structured boundary responses fall into the zero bin, yielding a high $\rho_0 = \mathbb{P}(|x_{\text{boundary}}| \leq \Delta/2)$. Once zeroed, subsequent attention mixing cannot recover the missing signed evidence. The collapse is activation-dominated (W4A8 stays close to FP32 while W4A4 fails on CFRN [42], as shown in Tab. 1), suggesting that COD needs token-local range control that explicitly bounds both the resolution ratio $\eta = \Delta/\sigma$ and the zero-bin mass $\rho_0$ (Fig. 3).

Based on diagnosis, we propose **COD-TDQ**, a COD-aware dynamic activation quantization framework for W4A4 Transformer-based COD. COD-TDQ remains purely PTQ (no retraining) and is hardware-agnostic. It combines Direct-Sum Token-Group (**DSTG**) to assign token-group activation scales and remove cross-token **range domination**, and Dual-Constraint Range Projection

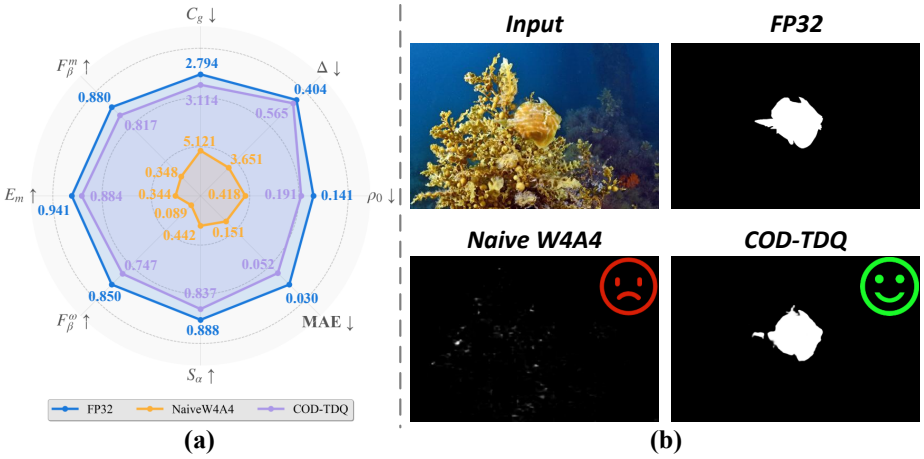


**Fig. 1: COD-specific W4A4 failure.** Naive W4A4 inflates a shared clipping range, producing a coarse step size and high zero-bin mass that erases weak boundary evidence. The inset summarizes representative diagnostics ($c_g, \Delta, \rho_0$) and the associated $S_\alpha$ collapse/recovery on CFRN/NC4K (rounded to three decimals).

**Table 1: Motivating observation** on NC4K (CFRN backbone).

| Setting | $S_\alpha \uparrow$ | $F_\beta^\omega \uparrow$ | $E_m \uparrow$ | $F_\beta^m \uparrow$ | MAE$\downarrow$ |
|---|---|---|---|---|---|
| FP32 (baseline) | **.888** | **.850** | **.941** | **.880** | **.030** |
| W8A8 (naive) | .887 | .850 | .940 | .879 | .030 |
| W4A8 (naive) | .882 | .841 | .936 | .872 | .032 |
| W4A4 (naive) | .443 | .089 | .344 | .348 | .151 |

(**DCRP**) to project each token-group range so that the step ratio $\eta$ and the **zero-bin mass** $\rho_0$ stay in a stable regime. Across four COD benchmarks and two Transformer COD models (CFRN and ESCNet [54]), we perform comprehensive and extensive evaluations of COD-TDQ. The results consistently show that COD-TDQ significantly surpasses representative PTQ baselines under the same W4A4 quantization protocol, delivering over 0.12–0.14 improvements in $S_\alpha$ (CFRN) compared with the state-of-the-art quantization approach without retraining.

Our contributions can be summarized as follows:

- We provide a mechanism-driven diagnosis of COD-specific W4A4 collapse, centered on **range domination** and **zero-bin mass** collapse, with diagnostic metrics of $\Delta$, $\eta$ and $\rho_0$, respectively.
- We propose **COD-TDQ**, which stabilizes W4A4 activation ranges via (i) Direct-Sum Token-Group scaling (DSTG) to suppress **range domination**,

and (ii) Dual-Constraint Range Projection (DCRP), which enforces **step-to-dispersion** and **zero-bin-mass** constraints to bound $\eta$ and $\rho_0$.
– We establish a unified W4A4 PTQ benchmark for COD across four datasets and provide diagnostic metrics (i.e., $\Delta$, $\eta$, $\rho_0$) that explain baseline failures, facilitating future research on COD quantization.

## 2 Related Work

**PTQ for Transformers and Segmentation Models.** We focus on W4A4 PTQ for Transformer-based COD and relate our work to recent advances in Transformer PTQ. PTQ aims to convert pretrained networks to low precision without retraining, typically combining weight quantization with activation range selection using a calibration set and local reconstruction [21,58]. For vision transformers(ViT) [6], PTQ4ViT [56] exemplifies block-level calibration and reconstruction to make projections quantizable. Subsequent methods such as RepQ-ViT [24] and FIMA-Q [50] further reduce the quantization-induced representation drift by improving feature fidelity. Optimization-based rounding (AdaRound [34]), block reconstruction with scheduling (BRECQ [11]), and calibration regularization (QDrop [47]) which minimize reconstruction error around the calibration distribution. PQ-SAM [27] and PTQ4SAM [30] study PTQ strategies tailored to the Segment Anything Model (SAM), emphasizing sensitive pathways in attention and normalization and the interaction between prompt encoders and mask decoders. But the COD setting differs: informative signals often appear as small-amplitude, structured boundary responses embedded in background-driven heavy tails. This shifts the bottleneck from global fidelity to preserving weak token-local cues under aggressive activation quantization.

**Token Sensitivity and Activation Heterogeneity.** Ultra-low-bit quantization is particularly sensitive to heavy-tailed activations and heterogeneous statistics. ORQ-ViT [14] explicitly handles outliers to prevent a small number of extreme values from dominating the clipping range, while NoisyQuant [28] introduces noise and perturbation modeling to better match non-Gaussian activation behaviors. SmoothQuant [51] re-parameterizes activations and weights to ease activation quantization by shifting difficulty to weights, and AHCPTQ [57] treats activation heterogeneity as a first-class issue in calibration. The post-GELU token-based dynamic bit-width assignment [17] is representative in exploiting token statistics to decide where additional precision is needed. Such methods provide useful insight that token distributions are highly non-uniform, yet their main lever is changing the bit-width. IGQ-ViT [33] and ADFQ-ViT [16] explore adaptive grouping strategies to refine quantization granularity beyond a single global scale. Despite their different mechanisms, most of these methods still operate with shared ranges at the layer and block level and evaluate primarily on classification. Under COD at W4A4, the critical failures are token-local: weak boundary evidence can collapse into the zero bin, which is not directly implied by outlier suppression or average reconstruction fidelity. In our setting,

the quantization budget is fixed (W4A4) under a standard quantization protocol, so the dominant requirement becomes stabilizing *4-bit activations* without relying on dynamic bit-width execution. Bit allocation alone does not prevent token-local inflation or guarantee that minority boundaries avoid zeroization.

**Architecture-aware error control.** Cross-domain PTQ designs highlight the importance of structural constraints: ARCQuant [32] tailors quantization to residual and attention structures on NVFP4 [1], and QuaRTZ [18] emphasizes controlling error accumulation and sparsity patterns. While these ideas motivate architecture-aware quantization, they are not formulated around dense prediction failures driven by token-local activation range selection. In COD under W4A4, cross-token range domination inflates the quantization step for most tokens, and zero-bin mass collapse erases weak-but-structured boundary cues. COD-TDQ addresses this gap with *token-local, channel-group-wise* activation quantization (DSTG) and a *dual-constraint* range projection (DCRP) that bounds both the step-to-dispersion ratio and the zero-bin mass.

## 3 W4A4 Fragility Diagnosis for COD

This section provides a diagnosis for why Transformer-based COD is unusually fragile under post-training W4A4. The collapse forms a coupled loop: background-dominated activation statistics inflate a shared range and step size, which increases the zero-bin mass and erases weak boundary cues that COD relies on.

### 3.1 Preliminaries

The failure mechanism is easiest to expose under a conventional tensor-wise activation quantizer that shares a single range across all tokens. This subsection defines the diagnostic variables used throughout the analysis.

Consider an activation tensor $X \in \mathbb{R}^{T\times C}$ from a single layer and a single input sample, where $T$ indexes tokens and $C$ indexes channels. Let $x_c$ denote an element. A symmetric $a$-bit uniform quantizer selects a tensor-wise clip radius $c > 0$ and uses the integer grid $q_{\min} = -2^{a-1}$ and $q_{\max} = 2^{a-1} - 1$. For W4A4 activations, $a = 4$ so $q_{\max} = 7$. The corresponding step size is $\Delta = c/q_{\max}$. Under rounding-to-nearest, the dequantized value satisfies $\hat{x}_c = 0$ if and only if $|x_c| \le \Delta/2$. Four scalar diagnostics summarize the key effects.

Global clip factor $c_g$. To compare clip ranges across layers with different scales, the clip radius is normalized by the tensor dispersion. Let $\sigma_g = \mathrm{Std}(X)+\varepsilon$ denote the standard deviation over all $TC$ elements, where $\varepsilon > 0$ is a small constant. The normalized global clip factor is

$$c_g \triangleq \frac{c}{\sigma_g}, \quad \Delta \triangleq \frac{c}{q_{\max}}, \quad \eta \triangleq \frac{\Delta}{\sigma_{\mathcal{A}}}. \tag{1}$$

Larger $c_g$ indicates that a wider range is allocated relative to the typical tensor.

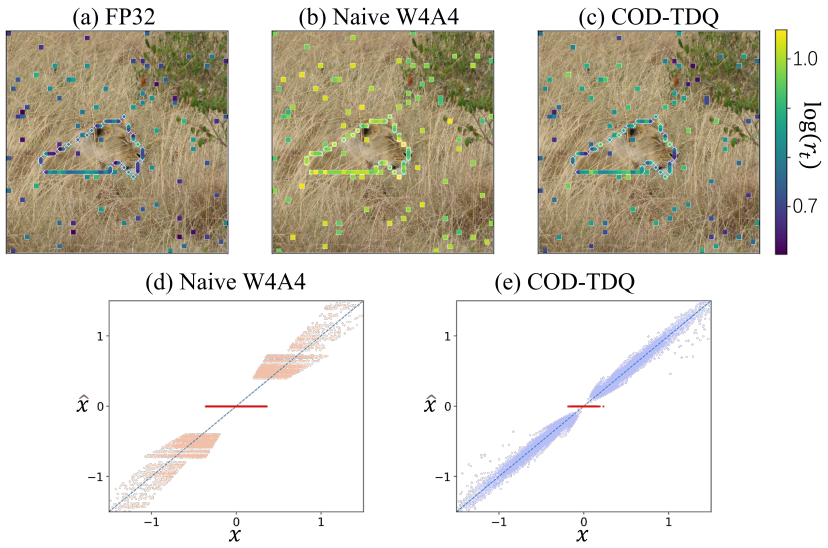


**Fig. 2: Reduces cross-token scale interference.** (a–c) Token-wise range disparity under FP32, naive W4A4, and DSTG: token-group scaling mitigates background-dominated range inflation. (d–e) Boundary-region activation magnitudes before/after quantization: naive W4A4 collapses many small responses to zero, while COD-TDQ preserves them, reducing the zeroed-activation fraction from 41.6% to 14.2%.

Step size $\Delta$. The step size is the quantization resolution within the unclipped region. For tensor-wise symmetric quantization, it is defined above, where $q_{\max}$ is the maximum representable quantized magnitude.

Step-to-dispersion ratio $\eta$. A small absolute step can still be coarse for weak cues. For a selected activation set $\mathcal{A}$ with dispersion $\sigma_{\mathcal{A}} = \mathrm{Std}(\mathcal{A}) + \varepsilon$, the resolution ratio is defined above. $\mathcal{A}$ is instantiated as boundary-heavy activations.

Zero-bin mass $\rho_0$. For a set of activations $\mathcal{A}$, the zero-bin mass is the fraction of elements that quantize to zero, where the probability view refers to the empirical distribution of the selected activations:

$$\rho_0 \triangleq \frac{1}{|\mathcal{A}|} \sum_{x \in \mathcal{A}} \mathbf{1}(\hat{x} = 0) = \mathbb{P}\left( |x| \leq \frac{\Delta}{2} \right). \tag{2}$$

### 3.2 Range Domination

COD is intentionally low-contrast, so useful evidence often takes the form of small-amplitude but spatially structured responses. At the same time, the token population is dominated by diverse backgrounds. This imbalance creates strong token-wise activation heterogeneity. A tensor-wise quantizer couples all tokens through a single clip radius $c$. A small number of heavy-tailed background spikes

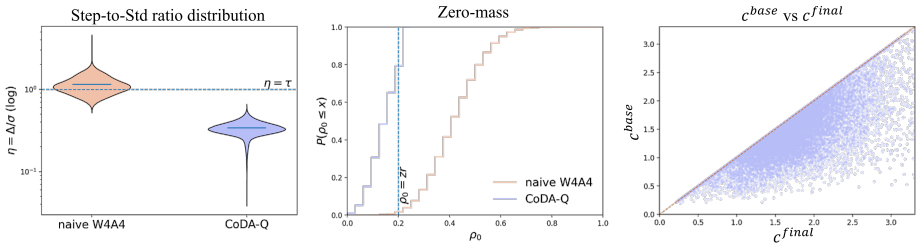


**Fig. 3: DCRP prevents zero-bin mass collapse.** DCRP projects each token-group clip radius to satisfy a step-to-dispersion bound and a zero-bin mass bound. The fraction of non-boundary token-groups exceeding the step-to-std threshold drops from 72.60% (pre-projection) to 0.00% after C1, and the fraction with pre-projection $\rho_0 > \mathrm{zr}$ drops from 98.36% (naive W4A4) to 20.87% under COD-TDQ statistics.

can therefore dominate the range selection, increase $c$, and enlarge the step size $\Delta$ in Eq. (1). Cross-token heterogeneity is summarized by the **range disparity**

$$\mathcal{D}(X) = \frac{\max_c |x_c|}{\mathrm{median}_t \ \mathrm{median}_c \ |x_c|}, \tag{3}$$

which becomes large when a small fraction of tokens carries extreme magnitudes. The shared range is set by outliers rather than by the majority of tokens, so most tokens are quantized with an overly coarse resolution. W4A4 amplifies the token-wise range disparity $\mathcal{D}(X)$ due to background spikes (Fig. 2), whereas token-group scaling (Sec. 4.2) markedly suppresses the cross-token range inflation.

### 3.3 Zero-bin Mass Collapse

Coarse step size becomes catastrophic in COD by increasing the zero-bin mass on boundary-heavy activations. Range domination becomes destructive in COD because the task depends on weak boundary cues. Once the global step size becomes coarse, many of these cues fall into the zero bin and disappear.

Under rounding-to-nearest, $\hat{x} = 0$ holds when $|x| \le \Delta/2$. The zero-bin mass in Eq. (2) therefore increases monotonically with $\Delta$ for any fixed activation distribution. For boundary-heavy activations, the increase is often steep because boundary responses cluster near zero. After zeroization, subsequent attention mixing and linear projections cannot recreate the missing signed evidence, since the input to those operations is exactly zero. This creates a token-local bottleneck that manifests as a global mask failure. Naive W4A4 places many token-groups in an unstable regime with overly large step-to-dispersion ratio $\eta$ and excessive zero-bin mass $\rho_0$, motivating explicit control of both diagnostics (Fig. 3).

**Evidence from Mechanism Diagnostics.** The diagnosis is linked to measurable forward-pass signals and to the representative numbers reported in Fig. 1. The coupled range-domination and zero-bin mass loop is measurable during forward passes. Fig. 1 reports representative diagnostics on CFRN evaluated on

NC4K. Naive W4A4 increases the failure signals together with the accuracy collapse. The bit-width diagnostic in Tab. 1 supports this view.

Naive W8A8 and W4A8 remain close to FP32, while naive W4A4 collapses sharply. Specifically, Tab. 1 shows that $S_\alpha$ drops from 0.888 (FP32) to 0.443 (naive W4A4), while the global clip factor increases from $c_g = 2.794$ to 5.121, the step size inflates from $\Delta = 0.404$ to 3.651, and the zero-bin mass rises from $\rho_0 = 0.141$ to 0.418. The same figure shows that COD-TDQ brings the diagnostics back toward the FP32 regime, with $S_\alpha = 0.837$, $c_g = 3.114$, $\Delta = 0.565$, and $\rho_0 = 0.191$ under W4A4, as shown in Fig. 1.

These measurements isolate two requirements for reliable W4A4 COD. First, range selection must be token-local to prevent background tokens from dominating the dynamic range. Second, the quantizer must explicitly control zeroization for boundary-heavy activations. Sec. 4 operationalizes these requirements with token-group ranges (DSTG) and dual-constraint range projection (DCRP). Detailed testable predictions and measurement protocols are provided in Supp. Sec. S2.5. The dominant failure factor is therefore 4-bit activations rather than 4-bit weights.

Static Symmetric Weight Quantization. Weights are quantized once with standard symmetric uniform quantization and remain fixed during inference. Activation robustness is dominated by the activation-side design in Sec. 4.2. Let $W \in \mathbb{R}^{O\times I}$ denote a weight matrix with output dimension $O$ and input dimension $I$. Let $s \in \mathbb{R}^O$ denote per-output-channel scales with $s_o > 0$. We use broadcasted division: $(W/s)_{o,i} = W_{o,i}/s_o$. Weight quantization is

$$Q_w(W) = \mathrm{clip}\left(\lfloor W/s \rceil,\ q^w_{\min}, q^w_{\max}\right), \qquad \hat{W} = s \odot Q_w(W), \tag{4}$$

where $\odot$ denotes element-wise multiplication with broadcasting along $I$. Packing and storage are implementation details and are summarized in Supp. Sec. S1.2.

## 4 Method

This section specifies post-training simulated quantization under W4A4 and details COD-TDQ, a token-local activation quantization framework for Transformer-based camouflaged object detection. The design follows the fragility diagnosis in Sec. 3 by suppressing cross-token range domination and by explicitly controlling the 4-bit step size and the induced zero-bin mass.

### 4.1 Overview of COD-TDQ

Guided by the fragility diagnosis in Sec. 3, COD-TDQ treats COD quantization stability as a token-local range selection problem, aiming to suppress cross-token **range domination** and to explicitly control the 4-bit step size and the induced **zero-bin mass**. To this end, COD-TDQ introduces two coupled modules—**DSTG** and **DCRP**—that operate at the token-group granularity and are detailed in Sec. 4.2 and Sec. 4.3, (Fig. 4). Forward pseudocode is given in the

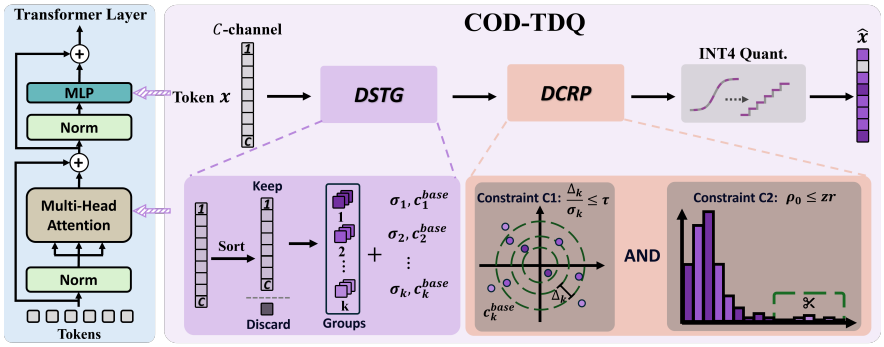


**Fig. 4: COD-TDQ (DSTG and DCRP) pipeline overview**.

supplements. Fig. 2(d–e) shows the practical payoff of this token-local design: compared to naive W4A4, COD-TDQ preserves many weak boundary activations that would otherwise be rounded into the zero bin.

Symmetric uniform quantization. The operator clip$(\cdot)$ clips element-wise to a closed interval. The notation $\lfloor \cdot \rceil$ denotes rounding to the nearest integer. Given a clip radius $c > 0$, the step size is $\Delta = c/q_{\max}$. A scalar $x$ is clipped to $\tilde{x} = \text{clip}(x, -c, c)$, quantized as $q = \text{clip}(\lfloor \tilde{x}/\Delta \rceil, q_{\min}, q_{\max})$, and dequantized as $\hat{x} = \Delta q$. A small constant $\varepsilon > 0$ avoids degenerate steps in implementation. COD-TDQ targets the dominant projection operators, namely `Linear` layers in attention and MLP blocks and `Conv` layers when present. Full operator coverage and implementation details are listed in Supp. Sec. S1.1.

COD-TDQ instantiates token-local range selection with two coupled modules. DSTG localizes scaling to token groups to remove cross-token range domination. DCRP then projects each group range so that discretization and zeroization remain bounded under 4-bit activations. **DSTG** (**D**irect-**S**um **T**oken-**G**roup) partitions each token vector into fixed-size channel groups and assigns each group its own activation range. **DCRP** (**D**ual-**C**onstraint **R**ange **P**rojection) adjusts each group range with two constraints that control the step-to-dispersion ratio and the zero-bin mass. We next detail DSTG in Sec. 4.2, and then introduce DCRP in Sec. 4.3 to complete the token-group W4A4 quantizer.

### 4.2 Direct-Sum Token-Group

Token-wise activation heterogeneity in COD lets heavy-tailed background tokens dominate this shared range, inflating the step size and increasing the zero-bin mass on boundary-heavy activations, as discussed in Sec. 3. This cross-token coupling motivates a token-local range assignment that decouples background-driven outliers from the majority of tokens. In response, DSTG assigns a dedicated activation range to each token-group, which prevents background tokens from dictating a shared clipping range. DSTG decomposes each token into groups and performs uniform quantization.

Direct-Sum Token-Group Decomposition. For a token vector $x \in \mathbb{R}^C$, channels are grouped into blocks of size $g$. If $C$ is not divisible by $g$, the vector is zero-padded on the channel dimension to $C_{\text{pad}} = g\lceil C/g \rceil$.

$$x = \bigoplus_{k=1}^{K} x_k, \qquad x_k \in \mathbb{R}^g, \qquad K = \frac{C_{\text{pad}}}{g}, \tag{5}$$

$$c_k^{\text{base}} = \begin{cases} \|x_k\|_\infty, & \text{without percentile clipping,} \\ Q_p\big(|x_k|\big), & \text{with percentile } p \in (0, 1], \end{cases} \tag{6}$$

The padded token is decomposed as Eq. (5). Where $\oplus$ denotes concatenation along the channel dimension. Padding is removed after dequantization. A base clip radius $c_k^{\text{base}} > 0$ is estimated from the magnitudes of the group vector $x_k$ as Eq. (6). Where $\|\cdot\|_\infty$ is the max-absolute norm and $Q_p$ is the empirical $p$-quantile over the $g$ elements of $|x_k|$. Token-group scaling plays an important role. Let $c^{\text{global}} = \max_k c_k^{\text{base}}$ denote a per-tensor radius used by a shared-range quantizer. Let $\sigma_k = \text{Std}(x_k) + \varepsilon$ denote the within-group standard deviation, computed over the $g$ elements. The corresponding step-to-dispersion ratio under a shared range satisfies

$$\frac{c^{\text{global}}}{q_{\max}\sigma_k} \gg \frac{c_k^{\text{base}}}{q_{\max}\sigma_k} \tag{7}$$

whenever $c^{\text{global}}$ is dominated by outliers from other tokens or groups. Such cross-token interference is frequent in COD and motivates token-group ranges.

Uniform signed quantization per token-group. Given the final clip radius $c_k$ after DCRP (Sec. 4.3), DSTG applies signed uniform quantization within the group:

$$\Delta_k = \max\left(\frac{c_k}{q_{\max}},\ \varepsilon\right), \qquad \tilde{x}_k = \text{clip}(x_k, -c_k, c_k), \tag{8}$$

$$q_k = \text{clip}\left(\left\lfloor \tilde{x}_k/\Delta_k \right\rceil,\ q_{\min}, q_{\max}\right) \in \mathbb{Z}^g, \qquad \hat{x}_k = \Delta_k\, q_k. \tag{9}$$

The quantized token is reconstructed as $\hat{x}_t = \bigoplus_k \hat{x}_k$, and the padded dimensions are removed to recover $C$ channels.

### 4.3 Dual-Constraint Range Projection

DSTG removes cross-token coupling, but a single token-group is still heavy-tailed and inflates its own range. DCRP therefore projects each group radius to satisfy two stability constraints that directly control discretization and zeroization.

**Constraint C1: step-to-dispersion bound.** The first constraint upper-bounds the step-to-dispersion ratio by a user-specified $\tau > 0$ ($\sigma_k = \text{Std}(x_k) + \varepsilon$).

$$\eta_k \triangleq \frac{\Delta_k}{\sigma_k} \le \tau \quad \Longleftrightarrow \quad c_k \le c_k^{(\tau)} \triangleq q_{\max}\tau\sigma_k. \tag{10}$$

**Constraint C2: zero-bin mass bound.** Under rounding-to-nearest, an element quantizes to zero if $|x| \le \Delta/2$. The empirical zero-bin mass is

$$\rho_{0,k} \triangleq \frac{1}{g}\sum_{i=1}^{g} \mathbf{1}\left(|x_{k,i}| \le \frac{\Delta_k}{2}\right), \quad c_k \le c_k^{(\mathrm{zr})} \triangleq 2q_{\max}\, Q_{\mathrm{zr}}\big(|x_k|\big). \tag{11}$$

A target bound $\mathrm{zr} \in (0,1)$ limits zeroization by enforcing $\rho_{0,k} \le \mathrm{zr}$. Let $Q_{\mathrm{zr}}(|x_k|)$ denote the empirical zr-quantile of the $g$ magnitudes in the group. The condition $\rho_{0,k} \le \mathrm{zr}$ is satisfied when $\Delta_k/2 \le Q_{\mathrm{zr}}(|x_k|)$, which yields the bound on $c_k$ shown above. The feasible interval for the clip radius and the corresponding projection are

$$\mathcal{C}_k = \Big(0,\ \min\{c_k^{(\tau)},\ c_k^{(\mathrm{zr})}\}\Big], \qquad c_k = \Pi_{\mathcal{C}_k}\big(c_k^{\mathrm{base}}\big) = \min\Big(c_k^{\mathrm{base}},\ c_k^{(\tau)},\ c_k^{(\mathrm{zr})}\Big). \tag{12}$$

We then set $c_k \leftarrow \max(c_k, \varepsilon)$. By construction, the projected radius satisfies

$$\eta_k \le \tau, \qquad \rho_{0,k} \le \mathrm{zr} \quad \text{(up to empirical-quantile discretization)}. \tag{13}$$

As illustrated in Fig. 3, the projection in Eq. (12) sharply reduces the fraction of token-groups violating the C1/C2 bounds on $(\eta_k, \rho_{0,k})$, preventing zero-bin mass collapse in practice. The clipping and rounding trade-off, together with rounding-noise bounds implied by Constraint C1, are provided in Supp. Sec. S2.3.

Putting the two modules together. DSTG removes cross-token range domination by assigning token-group ranges (Eq. (7)). DCRP prevents each group from drifting into an unstable W4A4 regime where the step is too coarse or where zeroization is excessive (Eq. (13)). This combination directly targets the failure loop analyzed in Sec. 3 and is validated by diagnostics in Sec. 5.4.

## 5 Experiments

### 5.1 Experimental setup

**Datasets.** We evaluate on four standard COD benchmarks: CAMO [19] (1000 train / 250 test), CHAMELEON [36] (76 images), COD10K [8] (test split with 2026 images), and NC4K [31] (4121 images). Unless stated otherwise, we report results on the official test splits using the original image resolutions and the evaluation protocols released by prior COD work. Following COD conventions [8], we report five metrics: $S_\alpha$ (structure measure), $F_\beta^\omega$ (weighted F-measure), $E_m$ (mean E-measure), $F_\beta^m$ (max F-measure), and MAE. For $S_\alpha, F_\beta^\omega, E_m, F_\beta^m$, higher is better, while lower MAE indicates better mask quality.

**Models.** We primarily study CFRN [42], a strong Swin-based COD model with Transformer encoder blocks and a COD-specific decoder. We also evaluate on ESCNet [54], a Pyramid Vision Transformer (PVT)-style Transformer COD model with edge/texture collaboration modules.

**Table 2: W4A4 post-training quantization results on CFRN.**

| Method | CAMO | | | | | CHAMELEON | | | | | COD10K | | | | | NC4K | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $S_\alpha$ | $F_\beta^\omega$ | $E_m$ | $F_\beta^m$ | MAE | $S_\alpha$ | $F_\beta^\omega$ | $E_m$ | $F_\beta^m$ | MAE | $S_\alpha$ | $F_\beta^\omega$ | $E_m$ | $F_\beta^m$ | MAE | $S_\alpha$ | $F_\beta^\omega$ | $E_m$ | $F_\beta^m$ | MAE |
| *Full Precision & Naive Quantization* | | | | | | | | | | | | | | | | | | | | |
| FP32 | .876 | .844 | .934 | .877 | .042 | .912 | .875 | .961 | .898 | .019 | .870 | .795 | .939 | .835 | .022 | .888 | .850 | .941 | .880 | .030 |
| Naive W8A8 | .875 | .843 | .933 | .876 | .042 | .912 | .875 | .961 | .897 | .019 | .869 | .795 | .938 | .834 | .022 | .887 | .850 | .940 | .879 | .030 |
| Naive W4A8 | .864 | .826 | .923 | .861 | .047 | .906 | .863 | .959 | .886 | .021 | .858 | .776 | .931 | .818 | .024 | .882 | .841 | .936 | .872 | .032 |
| Naive W4A4 | .418 | .061 | .314 | .332 | .182 | .447 | .083 | .351 | .316 | .141 | .471 | .071 | .376 | .242 | .093 | .443 | .089 | .344 | .348 | .151 |
| *Existing PTQ Methods* | | | | | | | | | | | | | | | | | | | | |
| NoisyQuant | .408 | .112 | .498 | .331 | .190 | .443 | .120 | .574 | .318 | .150 | .474 | .101 | .380 | .248 | .095 | .450 | .173 | .324 | .341 | .124 |
| PTQ4ViT | .410 | .061 | .352 | .318 | .181 | .440 | .080 | .380 | .321 | .140 | .479 | .070 | .393 | .252 | .091 | .456 | .081 | .340 | .355 | .153 |
| ORQ-ViT | .415 | .069 | .303 | .321 | .181 | .448 | .083 | .356 | .316 | .141 | .489 | .086 | .364 | .240 | .092 | .440 | .081 | .332 | .344 | .151 |
| post-GELU | .403 | .093 | .469 | .319 | .185 | .449 | .084 | .466 | .318 | .147 | .480 | .073 | .544 | .259 | .084 | .450 | .094 | .502 | .352 | .156 |
| SmoothQuant | .397 | .118 | .526 | .321 | .140 | .445 | .104 | .558 | .328 | .140 | .490 | .073 | .565 | .249 | .091 | .452 | .108 | .543 | .342 | .157 |
| PQ-SAM | .464 | .228 | .511 | .398 | .121 | .465 | .190 | .499 | .317 | .150 | .483 | .131 | .462 | .243 | .089 | .474 | .211 | .508 | .351 | .154 |
| QuaRTZ | .416 | .054 | .309 | .324 | .182 | .447 | .083 | .353 | .319 | .140 | .471 | .070 | .371 | .243 | .093 | .441 | .084 | .337 | .350 | .151 |
| AHCPTQ | .417 | .060 | .313 | .329 | .182 | .449 | .086 | .364 | .320 | .141 | .471 | .070 | .374 | .243 | .093 | .442 | .086 | .340 | .345 | .151 |
| ARCQuant | .534 | .510 | .657 | .629 | .117 | .593 | .464 | .619 | .572 | .096 | .640 | .455 | .676 | .570 | .079 | .683 | .569 | .711 | .674 | .097 |
| FIMA-Q | .595 | .549 | .659 | .553 | .130 | .603 | .476 | .682 | .551 | .096 | .611 | .452 | .698 | .518 | .076 | .663 | .572 | .733 | .629 | .089 |
| PTQ4SAM | .671 | .591 | .740 | .653 | .106 | .605 | .461 | .645 | .518 | .095 | .665 | .540 | .760 | .584 | .072 | .702 | .632 | .761 | .660 | .082 |
| IGQ-ViT | .674 | .601 | .748 | .658 | .113 | .692 | .607 | .763 | .659 | .069 | .670 | .538 | .778 | .592 | .069 | .710 | .651 | .776 | .688 | .080 |
| RepQ-ViT | .676 | .608 | .750 | .656 | .099 | .701 | .622 | .774 | .672 | .067 | .676 | .549 | .779 | .599 | .062 | .718 | .651 | .783 | .691 | .072 |
| **Ours** | **.813** | **.724** | **.864** | **.795** | **.070** | **.862** | **.758** | **.904** | **.823** | **.040** | **.802** | **.649** | **.864** | **.736** | **.038** | **.837** | **.747** | **.884** | **.817** | **.052** |

**Table 3: W4A4 post-training quantization on ESCNet.** We bold the best results and underline the second-best results, including ties, among all PTQ methods.

| Method | Baselines | | | Existing PTQ Methods | | | | | | | | | Ours |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | FP32 | W8A8 | W4A8 | ORQ | GELU | PQ | Noisy | PTQ4V | FIMA | PTQ4S | IGQ | RepQ | **COD-TDQ** |
| $S_\alpha \uparrow$ | .893 | .893 | .888 | .576 | .581 | .609 | .714 | .723 | .748 | .768 | .818 | .818 | **.881** |
| $F_\beta^\omega \uparrow$ | .864 | .864 | .858 | .368 | .375 | .429 | .651 | .658 | .662 | .701 | .727 | .751 | **.849** |
| $E_m \uparrow$ | .945 | .945 | .941 | .562 | .566 | .607 | .803 | .818 | .821 | .843 | .878 | .883 | **.936** |
| $F_\beta^m \uparrow$ | .887 | .887 | .883 | .459 | .467 | .526 | .684 | .700 | .710 | .741 | .790 | .791 | **.876** |
| MAE↓ | .028 | .028 | .029 | .120 | .119 | .111 | .093 | .080 | .078 | .062 | .055 | .052 | **.031** |

**Quantization protocol.** We focus on **W4A4**, and use W8A8/W4A8 only as diagnostic references to localize the failure source. Weights are quantized once with symmetric uniform quantization (Sec. 3.1). All comparisons are accuracy-centric under the same evaluation pipeline. In all experiments, we apply COD-TDQ to `Linear`/`Conv` operators, while quantizing `LayerNorm`/`Softmax` with a conventional FP16 routine. We use a single hyperparameter setting shared across all datasets and both CFRN/ESCNet baselines: $g = 32$, $\tau = 1.0$, and $\mathrm{zr} = 0.2$.

**Baselines and reproduction.** We compare COD-TDQ against representative Transformer PTQ methods and cross-domain PTQ transfers listed in Tab. 2 and Tab. 3. For baselines that require offline calibration, we use a 128-image calibration set. All results are obtained under the same evaluation codebase with identical preprocessing and post-processing as the original FP32 models. All quantization methods operate on the same set of quantized layers for fairness.

### 5.2 Main results on CFRN

Bit-width sensitivity supports the COD-specific failure diagnosis. Across all four datasets, *naive W4A8* stays close to FP32 (average $S_\alpha$: 0.8776 vs. 0.8864 for FP32), while *naive W4A4* collapses severely (average $S_\alpha$: 0.4448, average MAE:

**Table 4: Ablation studies on NC4K dataset.** Comparison of components on CFRN (left) and ESCNet (right) baselines. Best results are in **bold**.

| Method | CFRN Backbone | | | | |
|---|---|---|---|---|---|
| | $S_\alpha \uparrow$ | $F_\beta^\omega \uparrow$ | $E_m \uparrow$ | $F_\beta^m \uparrow$ | MAE↓ |
| Naive W4A4 | .443 | .089 | .344 | .348 | .151 |
| Per-tensor | .407 | .093 | .513 | .185 | .209 |
| DSTG only | .451 | .180 | .532 | .241 | .270 |
| DCRP only | .435 | .174 | .504 | .228 | .307 |
| **Ours** | **.837** | **.747** | **.884** | **.817** | **.052** |

| Method | ESCNet Backbone | | | | |
|---|---|---|---|---|---|
| | $S_\alpha \uparrow$ | $F_\beta^\omega \uparrow$ | $E_m \uparrow$ | $F_\beta^m \uparrow$ | MAE↓ |
| Naive W4A4 | .576 | .368 | .562 | .459 | .120 |
| Per-tensor | .597 | .378 | .548 | .448 | .111 |
| DSTG only | .617 | .416 | .579 | .488 | .106 |
| DCRP only | .771 | .684 | .805 | .744 | .062 |
| **Ours** | **.881** | **.849** | **.936** | **.876** | **.031** |

0.1417). This pattern localizes the dominant failure factor to 4-bit activations, consistent with Sec. 3. COD-TDQ achieves the best W4A4 accuracy on CFRN across all datasets (Tab. 2). Compared to the strongest W4A4 baselines, COD-TDQ improves $S_\alpha$ by **+11.8 to +16.1** points and reduces MAE by **0.008 to 0.020** absolute. On NC4K, COD-TDQ reaches $S_\alpha = 0.8365$ with MAE 0.0520, while the best baseline (RepQ-ViT [24]) attains $S_\alpha = 0.7182$ with MAE 0.0715.

PTQ4ViT [56] targets ViT quantization with twin uniform quantization and Hessian-guided scale selection, but its shared scale assumption is brittle under COD token heterogeneity. FIMA-Q [50] and RepQ-ViT [24] improve reconstruction fidelity, yet remain layer/block-centric and do not explicitly bound token-local $(\eta, \rho_0)$, leaving a consistent gap to COD-TDQ. Channel grouping or bit-width allocation. IGQ-ViT [33] alleviates channel-level outliers but does not remove cross-token interference, hence remains limited by token-local zeroization. post-GELU [17] assigns dynamic bit-widths, but without correcting token-wise scale mismatch or bounding $\rho_0$, it still fails under fixed W4A4. Outlier-centric methods. Outlier suppression or noise injection does not directly prevent boundary cues from collapsing into the zero bin when $\Delta$ is coarse (Tab. 4). Accordingly, ORQ-ViT [14] and NoisyQuant [28] remain far from FP32 on CFRN, and SmoothQuant [51] is mainly effective in higher-activation-bit regimes. Transferred PTQ methods target different activation pathologies and do not explicitly intervene on COD token-wise heterogeneity and boundary-sensitive zeroization. COD-TDQ differs by enforcing token-group constraints on Eq. (13).

### 5.3 Transferability to ESCNet

COD-TDQ generalizes across Transformer backbones. On ESCNet, naive W4A4 also degrades sharply. COD-TDQ restores ESCNet to near-lossless W4A4 performance. Full W4A4 tables on all datasets are in Supp. Sec. S3.3. RepQ-ViT and IGQ-ViT remain strong, but their channel-wise mechanisms do not prevent token-local zeroization. In contrast, COD-TDQ applies token-group constraints.

On CFRN, removing token-local scaling (Per-tensor) exposes severe cross-token range domination and does not recover W4A4 performance. DSTG only also fails when local clip radii are still inflated by tails, consistent with the zero-bin mass collapse diagnosis (Sec. 3). DCRP only in attention yields limited gains because unstable ranges in MLP projections can still erase boundary evidence.

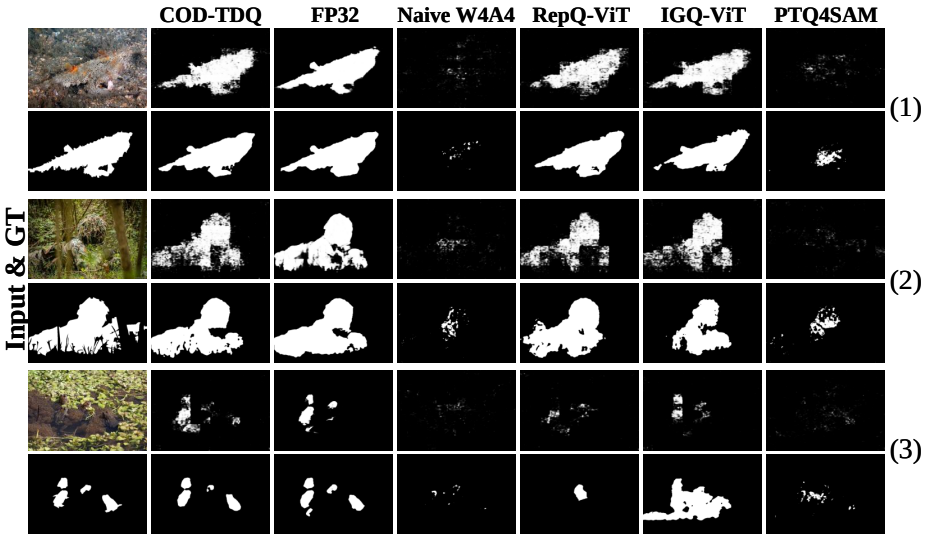


**Fig. 5: Qualitative comparison.** The first column shows the input image and the GT mask. The remaining columns present the prediction masks produced by different quantization methods (RepQ-ViT, IGQ-ViT, PTQ4SAM). For each example, the two rows correspond to results obtained with the CFRN and ESCNet baselines, respectively.

Only DSTG + DCRP consistently restores COD masks. On ESCNet, DCRP provides strong stabilization, and DSTG closes the remaining gap. On ESCNet, DCRP only already recovers a large portion of W4A4 accuracy, confirming that bounding $\Delta/\sigma$ and $\rho_0$ targets the dominant error mode. Adding DSTG further improves accuracy and reduces MAE, showing the complementarity of local scale alignment and constraint-based projection.

### 5.4 Qualitative Analysis

We visualize COD-TDQ's two components with mechanism-level diagnostics: DSTG reduces cross-token scale interference (Fig. 2) and DCRP enforces token-group stability bounds on $\eta = \Delta/\sigma$ and $\rho_0$ (Fig. 3).

Fig. 5 compares representative masks on challenging scenes (weak boundaries, textured backgrounds). Naive W4A4 often degenerates into nearly uniform predictions or fragmented responses. Strong ViT PTQ baselines (RepQ-ViT, IGQ-ViT) partially recover coarse structure but still miss fine contours. COD-TDQ produces masks closest to FP32, particularly on thin boundaries and low-contrast foreground regions.

### 5.5 Real INT4 Deployment

We further benchmark end-to-end CFRN inference using a Triton-based INT4 path on a single NVIDIA RTX 4090. We use $384 \times 384$ inputs and batch size

**Table 5:** Runtime efficiency comparison on CFRN.

| Method | Weight (MB)↓ | FPS (img/s)↑ | Latency (ms/img)↓ | Peak Memory (MB)↓ |
|---|---|---|---|---|
| FP32 | 775.40 | 29.40 | 34.02 | 1421.31 |
| Naive W4A4 | 108.30 | 43.95 | 22.75 | 843.30 |
| PTQ4SAM | **108.13** | 43.83 | 22.82 | 848.80 |
| RepQ-ViT | 108.15 | 44.18 | 22.64 | 844.67 |
| IGQ-ViT | 108.14 | 44.19 | 22.63 | 852.23 |
| **COD-TDQ** | 108.19 | **44.21** | **22.61** | **834.92** |

4. Latency is averaged per image, and memory denotes peak allocated GPU memory.

As shown in Tab. 5, COD-TDQ reaches 44.21 FPS and 22.61 ms/image, corresponding to a 1.50× throughput gain over FP32, while reducing the deployment artifact from 775.40 to 108.19 MB and peak memory from 1421.31 to 834.92 MB. Its runtime is also comparable to the other W4A4 methods. Relative to the implementation without DSTG (44.48 FPS), DSTG incurs only a 0.27 FPS (0.6%) overhead. The Triton path covers 69.73% of quantized operators. The ideal packed INT4 weights occupy 97,065,266 bytes. Scale tensors and indexing metadata add 326,632 and 4,364 bytes, respectively, giving only 0.341% combined storage overhead. DSTG uses contiguous token-group/channel-group scale indices and requires neither sorting nor an additional activation pass.

## 6 Conclusion

Camouflaged object detection (COD) attains high accuracy, yet Transformer-based COD models remain costly for mobile and edge deployment. Post-training quantization (PTQ) with 4-bit weights and activations (W4A4) is appealing. However, COD exhibits a pronounced, task-specific accuracy cliff. We trace this degradation mainly to 4-bit activations: token-wise heterogeneity and heavy-tailed background tokens dominate a shared clipping range, enlarge the step size, and suppress weak but structured boundary cues. To counter this mechanism without retraining, we propose COD-TDQ, integrating Direct-Sum Token-Group scaling (DSTG) with Dual-Constraint Range Projection (DCRP). COD-TDQ bounds both the discretization strength and zero-bin mass per token group. We anticipate this work will foster deployment-friendly COD quantization and inspire further study of tighter constraints, decoder diagnosis, cross-architecture generalization, and low-cost adaptation.

## Acknowledgements

This work is supported by the Natural Science Foundation of China (No. 62576176). The computational resources are supported by the Supercomputing Center of Nankai University (NKSC).

## References


1. Abecassis, F., Agrusa, A., Ahn, D., Alben, J., Alborghetti, S., Andersch, M., Arayandi, S., Bjorlin, A., Blakeman, A., Briones, E., et al.: Pretraining large language models with nvfp4. arXiv preprint arXiv:2509.25149 (2025)
2. Ariff, S.H.S., Liu, Y., Sun, G., Yang, J., Ding, H., Geng, X., Jiang, X.: Evaluating sam2 for video semantic segmentation. Machine Intelligence Research (2026)
3. Banner, R., Nahshan, Y., Soudry, D.: Post training 4-bit quantization of convolutional networks for rapid-deployment. NeurIPS **32** (2019)
4. Chen, X., Ren, G., Dai, T., Stathaki, T., Liu, H.: Enhancing prompt generation with adaptive refinement for camouflaged object detection. In: ICCV. pp. 20672–20682 (2025)
5. Das, B., Gopalakrishnan, V.: Camouflage anything: Learning to hide using controlled out-painting and representation engineering. In: CVPR. pp. 3603–3613 (2025)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
7. Du, J., Hao, F., Yu, M., Kong, D., Wu, J., Wang, B., Xu, J., Li, P.: Shift the lens: Environment-aware unsupervised camouflaged object detection. In: CVPR. pp. 19271–19282 (2025)
8. Fan, D.P., Ji, G.P., Sun, G., Cheng, M.M., Shen, J., Shao, L.: Camouflaged object detection. In: CVPR. pp. 2777–2787 (2020)
9. Frantar, E., Ashkboos, S., Hoefler, T., Alistarh, D.: Gptq: Accurate post-training quantization for generative pre-trained transformers. arXiv preprint arXiv:2210.17323 (2022)
10. Gao, Y., Kang, S., He, X., Li, B., Cheng, X., Liu, Y.: CATP: confidence-aware token pruning for camouflaged object detection. arXiv preprint arXiv:2604.16854 (2026)
11. Gong, R., Liu, X., Li, Y., Fan, Y., Wei, X., Guo, J.: Pushing the limit of post-training quantization. IEEE TPAMI (2025)
12. Guo, M.H., Lu, C.Z., Liu, Z.N., Cheng, M.M., Hu, S.M.: Visual attention network. Computational visual media **9**(4), 733–752 (2023)
13. Hao, C., Yu, Z., Liu, X., Xu, J., Yue, H., Yang, J.: A simple yet effective network based on vision transformer for camouflaged object and salient object detection. IEEE TIP (2025)
14. He, X., Lu, Y., Liu, H., Gong, C., He, W.: Orq-vit: Outlier resilient post training quantization for vision transformers via outlier decomposition. Journal of Systems Architecture p. 103530 (2025)
15. Ji, W., Li, J., Bi, Q., Liu, T., Li, W., Cheng, L.: Segment anything is not always perfect: An investigation of sam on different real-world applications. Machine Intelligence Research **21**, 617–630 (2024)
16. Jiang, Y., Sun, N., Xie, X., Yang, F., Li, T.: Adfq-vit: Activation-distribution-friendly post-training quantization for vision transformers. Neural Networks **186**, 107289 (2025)
17. Kim, D., Moon, J., Lee, J., Lee, G., Jeon, J., Ham, B.: Token-based dynamic bit-width assignment for vit quantization. PR p. 112269 (2025)
18. Kim, D., Lee, D., Chang, I.J., Bae, S.H.: Post-training quantization via residual truncation and zero suppression for diffusion models. arXiv preprint arXiv:2509.26436 (2025)

19. Le, T.N., Nguyen, T.V., Nie, Z., Tran, M.T., Sugimoto, A.: Anabranch network for camouflaged object segmentation. Computer vision and image understanding **184**, 45–56 (2019)
20. Lei, C., Fan, J., Li, X., Xiang, T.z., Li, A., Zhu, C., Zhang, L.: Towards real zero-shot camouflaged object segmentation without camouflaged annotations. IEEE TPAMI (2025)
21. Li, M., Zhang, F., Zhang, C.: Branch convolution quantization for object detection. Machine Intelligence Research **21**, 1192–1200 (2024)
22. Li, T., Guo, T., Xiang, D.: Lersgan: A gan-based model for low-light remote sensing image enhancement. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2025)
23. Li, Y.X., Chen, C.L.Z., Li, S., Hao, A.M., Qin, H.: A novel divide and conquer solution for long-term video salient object detection. Machine Intelligence Research **21**, 684–703 (2024)
24. Li, Z., Xiao, J., Yang, L., Gu, Q.: Repq-vit: Scale reparameterization for post-training quantization of vision transformers. In: ICCV. pp. 17227–17236 (2023)
25. Liu, J., Kong, L., Chen, G.: Improving sam for camouflaged object detection via dual stream adapters. In: ICCV. pp. 21906–21916 (2025)
26. Liu, J., Kong, L., Chen, G.: Improving sam for camouflaged object detection via dual stream adapters. In: ICCV. pp. 21906–21916 (2025)
27. Liu, X., Ding, X., Yu, L., Xi, Y., Li, W., Tu, Z., Hu, J., Chen, H., Yin, B., Xiong, Z.: Pq-sam: Post-training quantization for segment anything model. In: ECCV. pp. 420–437. Springer (2024)
28. Liu, Y., Yang, H., Dong, Z., Keutzer, K., Du, L., Zhang, S.: Noisyquant: Noisy bias-enhanced post-training activation quantization for vision transformers. In: CVPR. pp. 20321–20330 (2023)
29. Liu, Y., Wu, Y.H., Sun, G., Zhang, L., Chhatkuli, A., Gool, L.V.: Vision transformers with hierarchical attention. Machine Intelligence Research **21**, 670–683 (2024)
30. Lv, C., Chen, H., Guo, J., Ding, Y., Liu, X.: Ptq4sam: Post-training quantization for segment anything. In: CVPR. pp. 15941–15951 (2024)
31. Lv, Y., Zhang, J., Dai, Y., Li, A., Liu, B., Barnes, N., Fan, D.P.: Simultaneously localize, segment and rank the camouflaged objects. In: CVPR. pp. 11591–11601 (2021)
32. Meng, H., Luo, Y., Zhao, Y., Liu, W., Zhang, P., Ma, X.: Arcquant: Boosting nvfp4 quantization with augmented residual channels for llms. arXiv preprint arXiv:2601.07475 (2026)
33. Moon, J., Kim, D., Cheon, J., Ham, B.: Instance-aware group quantization for vision transformers. In: CVPR. pp. 16132–16141 (2024)
34. Nagel, M., Amjad, R.A., Van Baalen, M., Louizos, C., Blankevoort, T.: Up or down? adaptive rounding for post-training quantization. In: ICML. pp. 7197–7206. PMLR (2020)
35. Pang, Y., Zhao, X., Zuo, J., Zhang, L., Lu, H.: Open-vocabulary camouflaged object segmentation. In: ECCV. pp. 476–495. Springer (2024)
36. Portmann, A.: Animal camouflage. University of Michigan Press (1959)
37. Qian, Z., Li, T., Guo, S., Wang, B.: Dehazeswinunet: A swin transformer-based architecture for high-performance image dehazing. In: International Conference on Intelligent Computing. pp. 487–499. Springer (2025)
38. Ranjan, N., Savakis, A.: Lrp-qvit: Mixed-precision vision transformer quantization using layer importance score. In: 2025 25th International Conference on Digital Signal Processing (DSP). pp. 1–5. IEEE (2025)

39. Ren, G., Liu, H., Lazarou, M., Stathaki, T.: Multi-modal segment anything model for camouflaged scene segmentation. In: ICCV. pp. 19882–19892 (2025)
40. Ren, G., Liu, H., Lazarou, M., Stathaki, T.: Multi-modal segment anything model for camouflaged scene segmentation. In: ICCV. pp. 19882–19892 (2025)
41. Ren, P., Bai, T., Sun, F.: Esnet: An efficient skeleton-guided network for camouflaged object detection. Knowledge-Based Systems **311**, 113056 (2025)
42. Song, Z., Kang, X., Wei, X., Liu, J., Lin, Z., Li, S.: Continuous feature representation for camouflaged object detection. IEEE TIP (2025)
43. Sun, G., An, Z., Liu, Y., Liu, C., Sakaridis, C., Fan, D.P., Van Gool, L.: Indiscernible object counting in underwater scenes. In: CVPR. pp. 13791–13801 (2023)
44. Sun, K., Chen, Z., Lin, X., Sun, X., Liu, H., Ji, R.: Conditional diffusion models for camouflaged and salient object detection. IEEE TPAMI **47**(4), 2833–2848 (2025)
45. Sun, Y., Lian, J., Yang, J., Luo, L.: Controllable-lpmoe: Adapting to challenging object segmentation via dynamic local priors from mixture-of-experts. In: ICCV. pp. 22327–22337 (2025)
46. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. NeurIPS **30** (2017)
47. Wei, X., Gong, R., Li, Y., Liu, X., Yu, F.: Qdrop: Randomly dropping quantization for extremely low-bit post-training quantization. arXiv preprint arXiv:2203.05740 (2022)
48. Wu, D., Wang, M., Sun, J., Jia, X.: Knowledge-guided and collaborative learning network for camouflaged object detection. Engineering Applications of Artificial Intelligence **153**, 110771 (2025)
49. Wu, Y.H., Liu, W., Zhu, Z.X., Wang, Z., Liu, Y., Zhen, L.: GAPNet: A lightweight framework for image and video salient-object detection via granularity-aware paradigm. Machine Intelligence Research (2026)
50. Wu, Z., Wang, S., Zhang, J., Chen, J., Wang, Y.: Fima-q: Post-training quantization for vision transformers by fisher information matrix approximation. In: CVPR. pp. 14891–14900 (2025)
51. Xiao, G., Lin, J., Seznec, M., Wu, H., Demouth, J., Han, S.: Smoothquant: Accurate and efficient post-training quantization for large language models. In: ICML. pp. 38087–38099. PMLR (2023)
52. Xu, L., Xie, H., Qin, S.J., Tao, X., Wang, F.L.: Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. IEEE TPAMI (2026)
53. Yan, F., Jiang, X., Lu, Y., Cao, J., Chen, D., Xu, M.: Wavelet and prototype augmented query-based transformer for pixel-level surface defect detection. In: CVPR. pp. 23860–23869 (2025)
54. Ye, S., Chen, X., Zhang, Y., Lin, X., Cao, L.: Escnet: Edge-semantic collaborative network for camouflaged object detection. In: ICCV. pp. 20053–20063 (2025)
55. Yin, B., Zhang, X., Fan, D.P., Jiao, S., Cheng, M.M., Van Gool, L., Hou, Q.: Camoformer: Masked separable attention for camouflaged object detection. IEEE TPAMI **46**(12), 10362–10374 (2024)
56. Yuan, Z., Xue, C., Chen, Y., Wu, Q., Sun, G.: Ptq4vit: Post-training quantization for vision transformers with twin uniform quantization. In: ECCV. pp. 191–207. Springer (2022)
57. Zhang, W., Zhong, Y., Ando, S., Yoshioka, K.: Ahcptq: Accurate and hardware-compatible post-training quantization for segment anything model. In: CVPR. pp. 22383–22392 (2025)

58. Zhang, Z., Gao, Y., Fan, J., Zhao, Z., Yang, Y., Yan, S.: SelectQ: Calibration data selection for post-training quantization. Machine Intelligence Research **22**, 499–510 (2025)
59. Zhao, K., Yuan, W., Wang, Z., Li, G., Zhu, X., Fan, D.P., Zeng, D.: Open-vocabulary camouflaged object segmentation with cascaded vision language models. Computational Visual Media (2026)
60. Zhou, Y., Sun, G., Li, Y., Xie, G.S., Benini, L., Konukoglu, E.: When sam2 meets video camouflaged object segmentation: A comprehensive evaluation and adaptation. Visual Intelligence **3**(1), 10 (2025)
61. Zhou, Z., Li, Y., Zhong, C., Huang, J., Pei, J., Li, H., Tang, H.: Rethinking detecting salient and camouflaged objects in unconstrained scenes. In: ICCV. pp. 22372–22382 (2025)