当 W4A4 破坏伪装目标检测：Token 组双约束激活量化

李天淇[1], 方文昱[1], 何欣[2], 耿雪[3], 程徐[2], and 刘云[1,4,5⋆]

[1] 南开大学计算机学院，VCIP
[2] 天津理工大学计算机科学与工程学院
[3] 新加坡科技研究局
[4] 南开大学前沿交叉学科研究院
[5] 南开国际先进研究院（深圳福田）

Abstract. 伪装目标检测（Camouflaged Object Detection, COD）分割的是有意融入背景的目标，因此预测依赖细微的纹理与边界线索。COD 常常需要在严格的端侧内存与时延预算下运行，因此低比特推理具有很高需求。然而，COD 对激进量化异常敏感。我们研究基于 Transformer 的 COD 的训练后 W4A4 量化，并发现一个任务特有的性能断崖：重尾背景 token 主导共享激活范围，扩大步长，并将微弱但有结构的边界线索推入零桶。这暴露出一个 token 局部瓶颈——需要消除跨 token 的范围支配，并在 4-bit 激活下约束零桶质量。为此，我们提出 **COD-TDQ**，即一种 **COD** 感知的 **T**oken 组 **D**ual-constraint（双约束）激活 **Q**uantization（量化）方法。COD-TDQ 用两个耦合步骤解决该 token 局部瓶颈：**D**irect-**S**um **T**oken-**G**roup（**DSTG**）分配 *token* 组尺度以抑制跨 token 范围支配；**D**ual-**C**onstraint **R**ange **P**rojection（**DCRP**）则投影每个 token 组的截断范围，使步长-离散度比和零桶质量保持有界。在四个 COD 基准和两个基线模型（CFRN 与 ESCNet）上，COD-TDQ 在无需重训练的情况下，持续取得比最先进量化方法高出 0.12 以上的 S_α 分数。代码见 https://github.com/MCG-NKU/nku-model-compre。

Keywords: 伪装目标检测 · 训练后量化 · Token 组量化 · 双约束量化

1 引言

伪装目标检测（COD）通常被评估为对有意融入背景的目标进行二值掩码预测。COD 模型必须依赖细微的纹理与边界线索，因此有用证据常表现为小幅值但有结构的响应 [43]。随着 Transformer [6,29,46] 编码器和多阶段设计逐渐普及 [2,4,12,22,37,55,60], COD 模型已经取得显著发展 [7,15,20,23,25,35,39,40,49,59,61]。然而，伴随这种增长，COD 模型正在变得更加复杂，从而增加部署时的内存和时延成本 [5, 26, 44, 45, 53]。与此同时，COD 越来越被期望在严格的内存与时延约束下运行 [10, 13, 41, 48]，尤其是在移动端和边缘设备应用中。在这一背景下，降低模型部署期间的计算和存储成本变得十分必要。一种实用方案是应用训练后量化（PTQ），它能够在不需要重训练的情况下有效降低内存占用和激活成本 [3]。

⋆ 通讯作者：Yun Liu (liuyun@nankai.edu.cn)

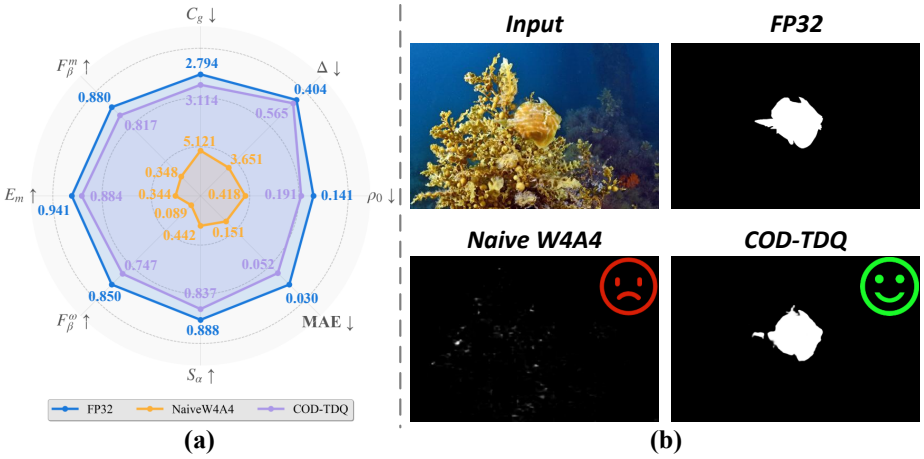

Fig. 1: COD 特有的 W4A4 失效。朴素 W4A4 会扩大共享截断范围，产生粗糙步长和较高零桶质量，从而擦除微弱边界证据。插图总结了代表性诊断量 (c_g, Δ, ρ_0) 以及 CFRN/NC4K 上对应的 S_α 坍塌/恢复（四舍五入到三位小数）。

在实践中，INT8 PTQ 通常是默认选择 [9, 52]。不过，当部署预算极其有限时，INT8 带来的收益可能有限，进一步降低成本需要进入超低比特设置。这促使我们在保持精度的同时探索 COD 的超低比特 PTQ。特别地，4-bit 权重与 4-bit 激活（W4A4）可在标准协议 [3,38] 下减少部署成本，同时不改变训练或推理流水线。然而，我们发现，基于 Transformer 的 COD 在朴素 W4A4 下可能发生坍塌，而不是平滑退化。这提出了一个关键问题：*是什么使 COD 在 4-bit 激活下如此脆弱，以及如何在不重训练、不依赖硬件特定假设的情况下使 W4A4 可靠？*

Transformer PTQ 发展迅速，但多数流水线仍落入几类反复出现的设计：(i) 在校准数据上进行逐块重构和舍入，以匹配 FP32 输出 [24, 50, 56]；(ii) 通过范围重参数化和平滑来缓解重尾激活 [51]；以及 (iii) 通过分组或自适应尺度获得更细粒度 [33]。这些策略大多以层或块为中心，并优化平均保真度。在 W4A4 的 COD 中，主导失效是 token 局部的：逐 token 激活异质性使背景 token 主导范围并增加零桶质量，而逐层拟合或重构并不会显式约束这一点。

我们将 W4A4 断崖追踪到**范围支配**与**零桶质量**耦合形成的坍塌机制（Fig. 1）。带有重尾激活尖峰的背景 token 会主导共享截断范围，扩大步长 Δ，并使大多数 token 的量化变粗（Fig. 1）。在最近邻舍入下，微弱但有结构的边界响应会落入零桶，产生较高的 $\rho_0 = \mathbb{P}(|x_{\text{boundary}}| \le \Delta/2)$。一旦被置零，后续注意力混合无法恢复缺失的有符号证据。这种坍塌由激活主导（W4A8 接近 FP32，而 W4A4 在 CFRN [42] 上失效，如 Tab. 1 所示），这表明 COD 需要 token 局部范围控制，并显式约束分辨率比 $\eta = \Delta/\sigma$ 与零桶质量 ρ_0（Fig. 3）。

基于上述诊断，我们提出 **COD-TDQ**，这是一个面向 W4A4 Transformer COD 的 COD 感知动态激活量化框架。COD-TDQ 保持纯 PTQ 形式（无需重训练），并且与硬件无关。它结合 Direct-Sum Token-Group（**DSTG**）来分配 token 组激活尺度并移除跨 token 的**范围支配**，以及 Dual-Constraint Range Projection（**DCRP**）来投影每个 token 组范围，使步长比 η 和**零桶质量** ρ_0 保

Table 1: NC4K（CFRN 骨干）上的**动机性观察**。

设置	$S_\alpha \uparrow$	$F_\beta^\omega \uparrow$	$E_m \uparrow$	$F_\beta^m \uparrow$	MAE↓
FP32（基线）	**.888**	**.850**	**.941**	**.880**	**.030**
W8A8（朴素）	.887	.850	.940	.879	.030
W4A8（朴素）	.882	.841	.936	.872	.032
W4A4（朴素）	.443	.089	.344	.348	.151

持在稳定区间。在四个 COD 基准以及两个 Transformer COD 模型（CFRN 与 ESCNet [54]）上，我们对 COD-TDQ 进行了全面而广泛的评估。结果持续表明，在相同 W4A4 量化协议下，COD-TDQ 显著超过代表性 PTQ 基线；在不重训练的情况下，与最先进量化方法相比，其 S_α（CFRN）提升超过 0.12–0.14。

我们的贡献总结如下：

- 我们提供了一个由机制驱动的 COD 特有 W4A4 坍塌诊断，该诊断以**范围支配**和**零桶质量**坍塌为核心，并分别使用 Δ、η 和 ρ_0 作为诊断指标。
- 我们提出 **COD-TDQ**，它通过 (i) Direct-Sum Token-Group 缩放（DSTG）抑制**范围支配**，并通过 (ii) Dual-Constraint Range Projection（DCRP）施加**步长-离散度**和**零桶质量**约束以约束 η 和 ρ_0，从而稳定 W4A4 激活范围。
- 我们为 COD 建立了跨四个数据集的统一 W4A4 PTQ 基准，并提供能够解释基线失效的诊断指标（即 Δ、η、ρ_0），以促进未来 COD 量化研究。

2 相关工作

Transformer 与分割模型的 PTQ。 我们聚焦于基于 Transformer 的 COD 的 W4A4 PTQ，并将本文工作与 Transformer PTQ 的近期进展联系起来。PTQ 旨在无需重训练地将预训练网络转换为低精度形式，通常结合权重量化、基于校准集的激活范围选择和局部重构 [21, 58]。对于视觉 Transformer（ViT） [6]，PTQ4ViT [56] 体现了块级校准与重构，以使投影可量化。RepQ-ViT [24] 和 FIMA-Q [50] 等后续方法通过提升特征保真度进一步降低量化引起的表征漂移。优化式舍入（AdaRound [34]）、带调度的块重构（BRECQ [11]）和校准正则化（QDrop [47]）则旨在最小化校准分布附近的重构误差。PQ-SAM [27] 和 PTQ4SAM [30] 研究面向 Segment Anything Model（SAM）的 PTQ 策略，强调注意力与归一化中的敏感路径，以及提示编码器与掩码解码器之间的交互。但 COD 场景不同：信息性信号常表现为小幅值、有结构的边界响应，并嵌入由背景驱动的重尾分布中。这将瓶颈从全局保真度转移到在激进激活量化下保留微弱的 token 局部线索。

Token 敏感性与激活异质性。 超低比特量化对重尾激活和异质统计尤其敏感。ORQ-ViT [14] 显式处理离群值，以防止少量极端值主导截断范围；NoisyQuant [28] 则引入噪声和扰动建模，以更好匹配非高斯激活行为。SmoothQuant [51] 通过将难度转移到权重上，对激活与权重进行重参数化以简化激活量化；AHCPTQ [57] 则把激活异质性视为校准中的一等问题。基于 post-GELU token 的动态比特宽

度分配 [17] 是利用 token 统计来决定何处需要额外精度的代表。这类方法提供了有用见解，即 token 分布高度非均匀，但其主要杠杆是改变比特宽度。IGQ-ViT [33] 和 ADFQ-ViT [16] 探索自适应分组策略，以在单一全局尺度之外细化量化粒度。尽管机制不同，这些方法大多仍在层级和块级使用共享范围，并主要在分类任务上评估。在 W4A4 的 COD 中，关键失效是 token 局部的：微弱边界证据可能坍塌到零桶，这并不能由离群值抑制或平均重构保真度直接推出。在我们的设置中，量化预算在标准量化协议下固定为 W4A4，因此主导需求是在不依赖动态比特宽度执行的情况下稳定 *4-bit 激活*。仅进行比特分配并不能防止 token 局部范围膨胀，也不能保证少数边界避免被置零。

架构感知的误差控制。 跨领域 PTQ 设计强调结构性约束的重要性：ARC-Quant [32] 针对 NVFP4 [1] 上的残差与注意力结构定制量化，QuaRTZ [18] 则强调控制误差累积和稀疏模式。尽管这些思想启发了架构感知量化，但它们并不是围绕由 token 局部激活范围选择驱动的密集预测失效来表述的。在 W4A4 的 COD 中，跨 token 范围支配会为大多数 token 扩大量化步长，而零桶质量坍塌会擦除微弱但有结构的边界线索。COD-TDQ 用 *token 局部*、*通道组级*激活量化（DSTG）以及*双约束*范围投影（DCRP）来弥补这一空白，同时约束步长-离散度比和零桶质量。

3 COD 的 W4A4 脆弱性诊断

本节诊断为什么基于 Transformer 的 COD 在训练后 W4A4 下异常脆弱。这种坍塌形成一个耦合回路：由背景主导的激活统计会膨胀共享范围和步长，从而增加零桶质量并擦除 COD 所依赖的微弱边界线索。

3.1 预备知识

在一个对所有 token 共享单一范围的传统逐张量激活量化器下，失效机制最容易被揭示。本小节定义贯穿分析所使用的诊断变量。

考虑来自单层和单个输入样本的激活张量 $X \in \mathbb{R}^{T\times C}$，其中 T 表示 token，C 表示通道。令 x_c 表示一个元素。对称 a-bit 均匀量化器选择逐张量截断半径 $c > 0$，并使用整数网格 $q_{\min} = -2^{a-1}$ 和 $q_{\max} = 2^{a-1} - 1$。对于 W4A4 激活，$a = 4$，因此 $q_{\max} = 7$。相应步长为 $\Delta = c/q_{\max}$。在最近邻舍入下，当且仅当 $|x_c| \le \Delta/2$ 时，反量化值满足 $\hat{x}_c = 0$。四个标量诊断量总结了关键影响。

全局截断因子 c_g。为了比较不同尺度层之间的截断范围，我们用张量离散度对截断半径进行归一化。令 $\sigma_g = \mathrm{Std}(X) + \varepsilon$ 表示所有 TC 个元素上的标准差，其中 $\varepsilon > 0$ 是一个小常数。归一化全局截断因子为

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

更大的 c_g 表示相对于典型张量分配了更宽的范围。

步长 Δ。步长是未截断区域内的量化分辨率。对于逐张量对称量化，其定义如上，其中 $q_{\max}$ 是可表示的最大量化幅值。

步长-离散度比 η。即使绝对步长很小，对微弱线索而言仍可能过粗。对于离散度为 $\sigma_{\mathcal{A}} = \mathrm{Std}(\mathcal{A}) + \varepsilon$ 的选定激活集合 $\mathcal{A}$，分辨率比如上定义。$\mathcal{A}$ 被实例化为边界密集激活。

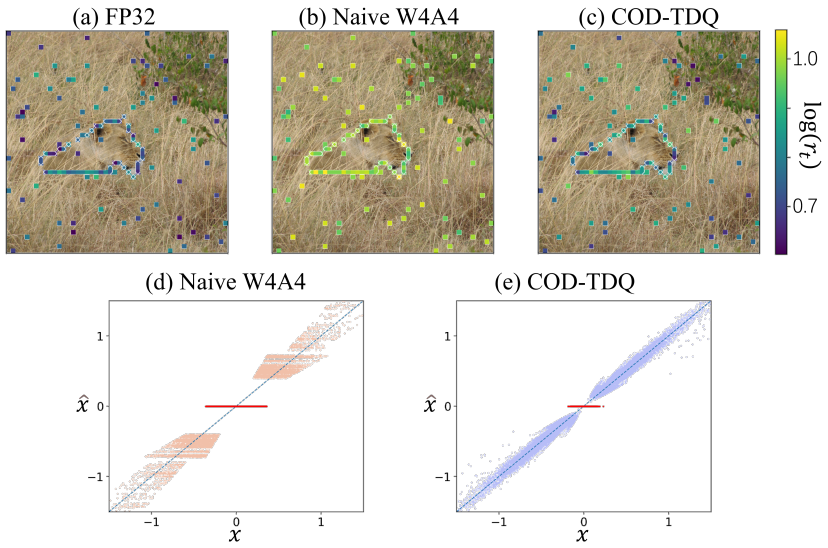

Fig. 2: 降低跨 token 尺度干扰。 (a–c) FP32、朴素 W4A4 与 DSTG 下的逐 token 范围差异：token 组缩放缓解了由背景主导的范围膨胀。(d–e) 量化前/后的边界区域激活幅值：朴素 W4A4 将许多小响应坍塌为零，而 COD-TDQ 保留了它们，将被置零激活比例从 41.6% 降至 14.2%。

零桶质量 ρ_0。对于激活集合 $\mathcal{A}$，零桶质量是被量化为零的元素比例；概率视角指的是选定激活的经验分布：

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

3.2 范围支配

COD 有意呈现低对比度，因此有用证据常以小幅值但空间结构化的响应形式出现。与此同时，token 群体由多样背景占据主导。这种不平衡造成强烈的逐 token 激活异质性。逐张量量化器通过单一截断半径 c 将所有 token 耦合在一起。因此，少量重尾背景尖峰可以主导范围选择，增大 c，并扩大 Eq. (1) 中的步长 Δ。跨 token 异质性由**范围差异**概括为

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

当一小部分 token 携带极端幅值时，该量会变大。共享范围由离群值而非大多数 token 决定，因此多数 token 被过粗的分辨率量化。由于背景尖峰，W4A4 会放大逐 token 范围差异 $\mathcal{D}(X)$（Fig. 2），而 token 组缩放（Sec. 4.2）显著抑制跨 token 范围膨胀。

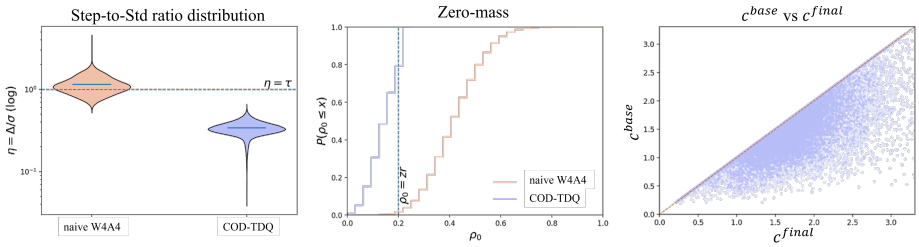

Fig. 3: DCRP 防止零桶质量坍塌。DCRP 将每个 token 组截断半径投影为同时满足步长-离散度界和零桶质量界。超过 step-to-std 阈值的非边界 token 组比例从 72.60%（投影前）在 C1 后降至 0.00%；预投影 $\rho_0 > \mathrm{zr}$ 的比例从 98.36%（朴素 W4A4）在 COD-TDQ 统计下降至 20.87%。

3.3 零桶质量坍塌

在边界密集激活上，粗糙步长会通过增加零桶质量而在 COD 中变得灾难性。范围支配在 COD 中具有破坏性，因为该任务依赖微弱边界线索。一旦全局步长变粗，许多这类线索就会落入零桶并消失。

在最近邻舍入下，当 $|x| \le \Delta/2$ 时有 $\hat{x} = 0$。因此，对于任意固定激活分布，Eq. (2) 中的零桶质量随 Δ 单调增加。对于边界密集激活，这种增加往往很陡，因为边界响应聚集在零附近。置零之后，后续注意力混合和线性投影无法重建缺失的有符号证据，因为这些操作的输入已经精确为零。这形成了一个 token 局部瓶颈，并表现为全局掩码失败。朴素 W4A4 使许多 token 组落入不稳定区间，其中步长-离散度比 η 过大、零桶质量 ρ_0 过高，这推动我们显式控制两个诊断量（Fig. 3）。

来自机制诊断的证据。 该诊断与可测量的前向传播信号以及 Fig. 1 中报告的代表性数值相关联。耦合的范围支配与零桶质量回路可以在前向传播期间测得。Fig. 1 报告了在 NC4K 上评估 CFRN 时的代表性诊断。朴素 W4A4 会在精度坍塌的同时增加失效信号。 Tab. 1 中的比特宽度诊断支持这一观点。

朴素 W8A8 与 W4A8 保持接近 FP32，而朴素 W4A4 急剧坍塌。具体而言，Tab. 1 表明 S_α 从 0.888（FP32）降至 0.443（朴素 W4A4），同时全局截断因子从 $c_g = 2.794$ 增至 5.121，步长从 $\Delta = 0.404$ 膨胀至 3.651，零桶质量从 $\rho_0 = 0.141$ 升至 0.418。同一图还表明 COD-TDQ 将诊断量拉回 FP32 区间附近：在 W4A4 下，$S_\alpha = 0.837$、$c_g = 3.114$、$\Delta = 0.565$、$\rho_0 = 0.191$，如 Fig. 1 所示。

这些测量分离出可靠 W4A4 COD 的两个要求。第一，范围选择必须是 token 局部的，以防止背景 token 主导动态范围。第二，量化器必须显式控制边界密集激活的置零。 Sec. 4 用 token 组范围（DSTG）和双约束范围投影（DCRP）实现这些要求。详细的可测试预测和测量协议见补充材料第 S2.5 节。因此，主导失效因素是 4-bit 激活，而不是 4-bit 权重。

静态对称权重量化。权重使用标准对称均匀量化一次性量化，并在推理期间保持固定。激活鲁棒性主要由 Sec. 4.2 中的激活侧设计决定。令 $W \in \mathbb{R}^{O\times I}$ 表示一个权重矩阵，其中输出维度为 O，输入维度为 I。令 $s \in \mathbb{R}^O$ 表示逐输出通道尺度，且 $s_o > 0$。我们使用广播除法：$(W/s)_{o,i} = W_{o,i}/s_o$。权重量化为

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

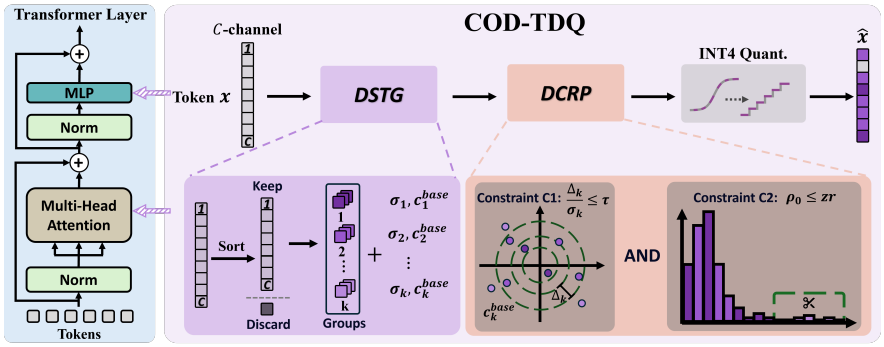

Fig. 4: COD-TDQ（DSTG 与 DCRP）流水线概览。

其中 $\odot$ 表示沿 I 广播的逐元素乘法。打包与存储是实现细节，并在补充材料第 S1.2 节中总结。

4 方法

本节规定 W4A4 下的训练后模拟量化，并详述 COD-TDQ——一个面向基于 Transformer 的伪装目标检测的 token 局部激活量化框架。该设计遵循 Sec. 3 中的脆弱性诊断，通过抑制跨 token 范围支配，并显式控制 4-bit 步长及其诱导的零桶质量。

4.1 COD-TDQ 概述

在 Sec. 3 中的脆弱性诊断指导下，COD-TDQ 将 COD 量化稳定性视为一个 token 局部范围选择问题，目标是抑制跨 token 的**范围支配**，并显式控制 4-bit 步长及其诱导的**零桶质量**。为此，COD-TDQ 引入两个耦合模块——**DSTG** 和 **DCRP**——它们在 token 组粒度上运行，并在 Sec. 4.2 和 Sec. 4.3 中详细说明（Fig. 4）。前向伪代码见补充材料。Fig. 2(d–e) 展示了这种 token 局部设计的实际收益：与朴素 W4A4 相比，COD-TDQ 保留了许多原本会被舍入到零桶中的微弱边界激活。

对称均匀量化。算子 $\mathrm{clip}(\cdot)$ 将元素逐个截断到闭区间。记号 $\lfloor\cdot\rceil$ 表示舍入到最近整数。给定截断半径 $c > 0$，步长为 $\Delta = c/q_{\max}$。标量 x 被截断为 $\tilde{x} = \mathrm{clip}(x, -c, c)$，量化为 $q = \mathrm{clip}(\lfloor \tilde{x}/\Delta \rceil, q_{\min}, q_{\max})$，并反量化为 $\hat{x} = \Delta q$。小常数 $\varepsilon > 0$ 用于避免实现中的退化步长。COD-TDQ 作用于主导投影算子，即注意力与 MLP 块中的 `Linear` 层，以及存在时的 `Conv` 层。完整算子覆盖范围和实现细节列于补充材料第 S1.1 节。

COD-TDQ 用两个耦合模块实例化 token 局部范围选择。DSTG 将缩放局部化到 token 组，以移除跨 token 范围支配。随后 DCRP 投影每个组范围，使 4-bit 激活下的离散化和置零保持有界。**DSTG**（**D**irect-**S**um **T**oken-**G**roup）将每个 token 向量划分为固定大小的通道组，并为每个组分配自身的激活范围。**DCRP**（**D**ual-**C**onstraint **R**ange **P**rojection）用两个约束调整每个组范围，以控制步

长-离散度比和零桶质量。接下来我们在 Sec. 4.2 中详述 DSTG，然后在 Sec. 4.3 中引入 DCRP，以完成 token 组 W4A4 量化器。

4.2 Direct-Sum Token-Group

如 Sec. 3 所讨论，COD 中逐 token 激活异质性使重尾背景 token 能够主导共享范围，扩大步长，并增加边界密集激活上的零桶质量。这种跨 token 耦合推动了 token 局部范围分配，使由背景驱动的离群值与大多数 token 解耦。作为回应，DSTG 为每个 token 组分配专用激活范围，从而防止背景 token 决定共享截断范围。DSTG 将每个 token 分解为若干组并执行均匀量化。

Direct-Sum Token-Group 分解。对于 token 向量 $x \in \mathbb{R}^C$，通道被分为大小为 g 的块。如果 C 不能被 g 整除，则在通道维度上将向量零填充到 $C_{\text{pad}} = g\lceil C/g \rceil$。

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

$$c_k^{\text{base}} = \begin{cases} \|x_k\|_\infty, & \text{不使用百分位截断,} \\ Q_p\big(|x_k|\big), & \text{使用百分位} p \in (0,1], \end{cases} \tag{6}$$

填充后的 token 按 Eq. (5) 分解。其中 $\oplus$ 表示沿通道维度拼接。反量化后移除填充维度。如 Eq. (6) 所示，基础截断半径 $c_k^{\text{base}} > 0$ 由组向量 x_k 的幅值估计。其中 $\|\cdot\|_\infty$ 是最大绝对范数，Q_p 是 $|x_k|$ 的 g 个元素上的经验 p 分位数。Token 组缩放起着重要作用。令 $c^{\text{global}} = \max_k c_k^{\text{base}}$ 表示共享范围量化器使用的逐张量半径。令 $\sigma_k = \text{Std}(x_k) + \varepsilon$ 表示组内标准差，它在 g 个元素上计算。在共享范围下，相应的步长-离散度比满足

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

只要 c^{global} 被其他 token 或组中的离群值主导。这种跨 token 干扰在 COD 中很常见，并推动使用 token 组范围。

每个 token 组的均匀有符号量化。给定 DCRP（Sec. 4.3）后的最终截断半径 c_k，DSTG 在组内应用有符号均匀量化：

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

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

量化后的 token 重构为 $\hat{x}_t = \bigoplus_k \hat{x}_k$，并移除填充维度以恢复 C 个通道。

4.3 双约束范围投影

DSTG 移除了跨 token 耦合，但单个 token 组仍可能重尾并膨胀自身范围。因此，DCRP 将每个组半径投影为满足两个稳定性约束，从而直接控制离散化和置零。

约束 C1：步长-离散度界。第一个约束将步长-离散度比上界限制为用户指定的 $\tau > 0$（$\sigma_k = \mathrm{Std}(x_k) + \varepsilon$）。

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

约束 C2：零桶质量界。在最近邻舍入下，如果 $|x| \le \Delta/2$，则元素被量化为零。经验零桶质量为

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

目标界 $\mathrm{zr} \in (0,1)$ 通过强制 $\rho_{0,k} \le \mathrm{zr}$ 来限制置零。令 $Q_{\mathrm{zr}}(|x_k|)$ 表示该组中 g 个幅值的经验 zr 分位数。当 $\Delta_k/2 \le Q_{\mathrm{zr}}(|x_k|)$ 时，条件 $\rho_{0,k} \le \mathrm{zr}$ 得到满足，由此得到上式所示的 c_k 界。截断半径的可行区间及其对应投影为

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

然后设置 $c_k \leftarrow \max(c_k, \varepsilon)$。由构造可知，投影后的半径满足

$$\eta_k \le \tau, \qquad \rho_{0,k} \le \mathrm{zr} \quad \text{（忽略经验分位数离散化误差）} \tag{13}$$

如 Fig. 3 所示， Eq. (12) 中的投影显著降低了违反 C1/C2 在 $(\eta_k, \rho_{0,k})$ 上界的 token 组比例，并在实践中防止零桶质量坍塌。截断与舍入之间的权衡，以及由约束 C1 推导的舍入噪声界，见补充材料第 S2.3 节。

将两个模块结合起来。DSTG 通过分配 token 组范围（Eq. (7)）移除跨 token 范围支配。DCRP 防止每个组漂移到步长过粗或置零过度的不稳定 W4A4 区间（Eq. (13)）。这种组合直接针对 Sec. 3 中分析的失效回路，并由 Sec. 5.4 中的诊断验证。

5 实验

5.1 实验设置

数据集。 我们在四个标准 COD 基准上评估：CAMO [19]（1000 张训练 / 250 张测试）、CHAMELEON [36]（76 张图像）、COD10K [8]（包含 2026 张图像的测试划分）以及 NC4K [31]（4121 张图像）。除非另有说明，我们使用原始图像分辨率和先前 COD 工作发布的评估协议，在官方测试划分上报告结果。按照 COD 惯例 [8]，我们报告五个指标：S_α（结构度量）、F_β^ω（加权 F 度量）、E_m（平均 E 度量）、F_β^m（最大 F 度量）以及 MAE。对于 $S_\alpha, F_\beta^\omega, E_m, F_\beta^m$，越高越好；MAE 越低表示掩码质量越好。

模型。 我们主要研究 CFRN [42]，这是一个强大的基于 Swin 的 COD 模型，包含 Transformer 编码器块和 COD 特定解码器。我们还在 ESCNet [54] 上评估，它是一个 PVT 风格的 Transformer COD 模型，具有边缘/纹理协作模块。

量化协议。 我们聚焦于 **W4A4**，并仅将 W8A8/W4A8 作为诊断参考来定位失效来源。权重使用对称均匀量化一次性量化（Sec. 3.1）。所有比较均在相同评估流

Table 2: CFRN 上的 **W4A4 训练后量化结果**。

方法	CAMO					CHAMELEON					COD10K					NC4K				
	S_α	F_β^ω	E_m	F_β^m	MAE	S_α	F_β^ω	E_m	F_β^m	MAE	S_α	F_β^ω	E_m	F_β^m	MAE	S_α	F_β^ω	E_m	F_β^m	MAE
全精度与朴素量化																				
FP32	.876	.844	.934	.877	.042	.912	.875	.961	.898	.019	.870	.795	.939	.835	.022	.888	.850	.941	.880	.030
朴素 W8A8	.875	.843	.933	.876	.042	.912	.875	.961	.897	.019	.869	.795	.938	.834	.022	.887	.850	.940	.879	.030
朴素 W4A8	.864	.826	.923	.861	.047	.906	.863	.959	.886	.021	.858	.776	.931	.818	.024	.882	.841	.936	.872	.032
朴素 W4A4	.418	.061	.314	.332	.182	.447	.083	.351	.316	.141	.471	.071	.376	.242	.093	.443	.089	.344	.348	.151
现有 *PTQ* 方法																				
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
本文方法	**.813**	**.724**	**.864**	**.795**	**.070**	**.862**	**.758**	**.904**	**.823**	**.040**	**.802**	**.649**	**.864**	**.736**	**.038**	**.837**	**.747**	**.884**	**.817**	**.052**

Table 3: ESCNet 上的 W4A4 训练后量化。在所有 PTQ 方法中，包括并列情形，我们将最优结果加粗，并将次优结果加下划线。

方法	基线			现有 PTQ 方法									本文方法
	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**

水线下以精度为中心。在所有实验中，我们将 COD-TDQ 应用于 `Linear`/`Conv` 算子，同时用传统 FP16 例程量化 `LayerNorm`/`Softmax`。我们在所有数据集以及 CFRN/ESCNet 两个基线上使用一组共享超参数：$g = 32$、$\tau = 1.0$ 和 $\mathrm{zr} = 0.2$。

基线与复现。 我们将 COD-TDQ 与 Tab. 2 和 Tab. 3 中列出的代表性 Transformer PTQ 方法及跨领域 PTQ 迁移方法进行比较。对于需要离线校准的基线，我们使用 128 张图像的校准集。所有结果均在相同评估代码库下获得，并采用与原始 FP32 模型相同的预处理和后处理。为保证公平，所有量化方法作用于同一组被量化层。

5.2 CFRN 上的主要结果

比特宽度敏感性支持 COD 特有的失效诊断。在全部四个数据集上，朴素 *W4A8* 保持接近 FP32（平均 S_α：0.8776，FP32 为 0.8864），而朴素 *W4A4* 严重坍塌（平均 S_α：0.4448，平均 MAE：0.1417）。该模式将主导失效因素定位到 4-bit 激活，与 Sec. 3 一致。COD-TDQ 在 CFRN 的所有数据集上取得最优 W4A4 精度（Tab. 2）。与最强 W4A4 基线相比，COD-TDQ 将 S_α 提升 **+11.8 到 +16.1** 个百分点，并将 MAE 的绝对值降低 **0.008 到 0.020**。在 NC4K 上，COD-TDQ 达到 $S_\alpha = 0.8365$、MAE 为 0.0520，而最佳基线（RepQ-ViT [24]）为 $S_\alpha = 0.7182$、MAE 为 0.0715。

Table 4: NC4K 数据集上的消融研究。在 CFRN（左）和 ESCNet（右）基线上比较各组件。最优结果以**粗体**标出。

方法	CFRN 骨干				
	$S_\alpha \uparrow$	$F_\beta^\omega \uparrow$	$E_m \uparrow$	$F_\beta^m \uparrow$	MAE↓
朴素 W4A4	.443	.089	.344	.348	.151
逐张量	.407	.093	.513	.185	.209
仅 DSTG	.451	.180	.532	.241	.270
仅 DCRP	.435	.174	.504	.228	.307
本文方法	**.837**	**.747**	**.884**	**.817**	**.052**

方法	ESCNet 骨干				
	$S_\alpha \uparrow$	$F_\beta^\omega \uparrow$	$E_m \uparrow$	$F_\beta^m \uparrow$	MAE↓
朴素 W4A4	.576	.368	.562	.459	.120
逐张量	.597	.378	.548	.448	.111
仅 DSTG	.617	.416	.579	.488	.106
仅 DCRP	.771	.684	.805	.744	.062
本文方法	**.881**	**.849**	**.936**	**.876**	**.031**

PTQ4ViT [56] 以双均匀量化和 Hessian 引导的尺度选择为目标处理 ViT 量化，但其共享尺度假设在 COD token 异质性下较脆弱。FIMA-Q [50] 和 RepQ-ViT [24] 提升了重构保真度，但仍以层/块为中心，并未显式约束 token 局部的 (η, ρ_0)，因此与 COD-TDQ 之间存在持续差距。通道分组或比特宽度分配。IGQ-ViT [33] 缓解了通道级离群值，但没有移除跨 token 干扰，因此仍受限于 token 局部置零。post-GELU [17] 分配动态比特宽度，但在未校正逐 token 尺度不匹配或约束 ρ_0 的情况下，在固定 W4A4 下仍会失效。以离群值为中心的方法。当 Δ 过粗时，离群值抑制或噪声注入并不能直接防止边界线索坍塌到零桶中（Tab. 4）。相应地，ORQ-ViT [14] 和 NoisyQuant [28] 在 CFRN 上仍远离 FP32，而 SmoothQuant [51] 主要在更高激活比特设置下有效。迁移而来的 PTQ 方法针对不同的激活病理，并没有显式干预 COD 的逐 token 异质性和边界敏感置零。COD-TDQ 的不同之处在于，它在 Eq. (13) 上施加 token 组约束。

5.3 向 ESCNet 的可迁移性

COD-TDQ 能跨 Transformer 骨干泛化。在 ESCNet 上，朴素 W4A4 也会急剧退化。COD-TDQ 将 ESCNet 恢复到近乎无损的 W4A4 性能。所有数据集上的完整 W4A4 表格见补充材料第 S3.3 节。RepQ-ViT 与 IGQ-ViT 仍然较强，但它们的逐通道机制无法防止 token 局部置零。相比之下，COD-TDQ 应用 token 组约束。

在 CFRN 上，移除 token 局部缩放（逐张量）会暴露严重的跨 token 范围支配，并无法恢复 W4A4 性能。当局部截断半径仍被尾部膨胀时，仅使用 DSTG 也会失败，这与零桶质量坍塌诊断一致（Sec. 3）。仅在注意力中使用 DCRP 只能带来有限收益，因为 MLP 投影中的不稳定范围仍可能擦除边界证据。只有 DSTG + DCRP 能够稳定恢复 COD 掩码。在 ESCNet 上，DCRP 提供强稳定作用，而 DSTG 弥合剩余差距。在 ESCNet 上，仅 DCRP 已经恢复了 W4A4 精度的很大一部分，确认约束 Δ/σ 和 ρ_0 针对的是主导错误模式。加入 DSTG 进一步提高精度并降低 MAE，表明局部尺度对齐和基于约束的投影具有互补性。

5.4 定性分析

我们用机制级诊断可视化 COD-TDQ 的两个组件：DSTG 降低跨 token 尺度干扰（Fig. 2），DCRP 则强制 token 组稳定性界 $\eta = \Delta/\sigma$ 和 ρ_0（Fig. 3）。

Fig. 5 比较了挑战性场景（弱边界、纹理背景）上的代表性掩码。朴素 W4A4 常退化为近乎均匀的预测或碎片化响应。较强的 ViT PTQ 基线（RepQ-ViT、

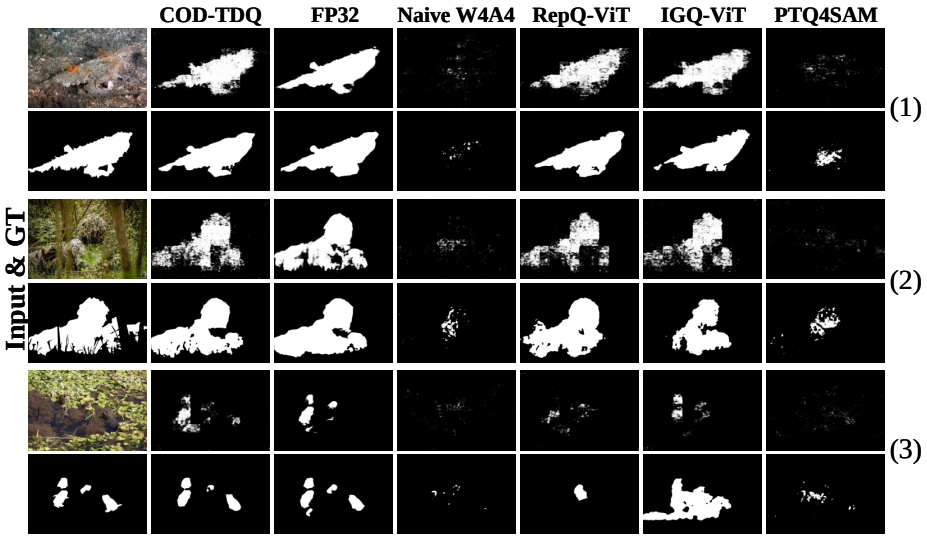

Fig. 5: 定性比较。第一列显示输入图像和 GT 掩码。其余列展示不同量化方法（RepQ-ViT、IGQ-ViT、PTQ4SAM）产生的预测掩码。对每个示例，两行分别对应 CFRN 和 ESCNet 基线得到的结果。

IGQ-ViT）可以部分恢复粗略结构，但仍会遗漏精细轮廓。COD-TDQ 产生最接近 FP32 的掩码，尤其是在细边界和低对比度前景区域。

5.5 真实 INT4 部署

我们进一步在单张 NVIDIA RTX 4090 上使用基于 Triton 的 INT4 路径，对 CFRN 端到端推理进行基准测试。我们使用 384×384 输入和 batch size 4。时延按每张图像平均，内存表示峰值分配 GPU 内存。

如 Tab. 5 所示，COD-TDQ 达到 44.21 FPS 和 22.61 ms/image，相当于较 FP32 获得 1.50× 吞吐增益，同时将部署产物从 775.40 MB 降至 108.19 MB，将峰值内存从 1421.31 MB 降至 834.92 MB。其运行时间也与其他 W4A4 方法相当。相对于不使用 DSTG 的实现（44.48 FPS），DSTG 仅带来 0.27 FPS（0.6%）开销。Triton 路径覆盖 69.73% 的量化算子。理想打包 INT4 权重占 97,065,266 字节。尺度张量和索引元数据分别增加 326,632 和 4,364 字节，总存储开销仅 0.341%。DSTG 使用连续 token 组/通道组尺度索引，既不需要排序，也不需要额外激活传递。

6 结论

伪装目标检测（COD）能够达到高精度，但基于 Transformer 的 COD 模型在移动端和边缘部署上仍然代价高昂。采用 4-bit 权重和激活（W4A4）的训练后量化（PTQ）具有吸引力。然而，COD 表现出明显且任务特有的精度断崖。我们将这种退化主要追踪到 4-bit 激活：逐 token 异质性和重尾背景 token 主导共

Table 5: CFRN 上的运行效率比较。

方法	权重 (MB)↓	FPS (img/s)↑	时延 (ms/img)↓	峰值内存 (MB)↓
FP32	775.40	29.40	34.02	1421.31
朴素 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**

享截断范围，扩大步长，并抑制微弱但有结构的边界线索。为在不重训练的情况下对抗这一机制，我们提出 COD-TDQ，它结合 Direct-Sum Token-Group 缩放（DSTG）与 Dual-Constraint Range Projection（DCRP）。COD-TDQ 对每个 token 组同时约束离散化强度和零桶质量。我们预期这项工作将促进部署友好的 COD 量化，并启发对更紧约束、解码器诊断、跨架构泛化和低成本适配的进一步研究。

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 International Conference on Digital Signal Processing (DSP). pp. 1–5 (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)