Zero-Shot Quantization for Object Detectors using Off-the-Shelf Generative Models

Hyunho Lee[1,*], Kyomin Hwang[1,*], Hyeonjin Kim[1,*], Suyoung Kim[1], Sunghyun Wee[1,2], and Nojun Kwak[1,†]

[1]Seoul National University, Seoul, South Korea
[2]LG Electronics, Seoul, South Korea

{hhlee822, kyomin98, peaceful1, ksyo96, wsh05, nojunk}@snu.ac.kr

Abstract. With an increasing number of Object Detection (OD) models being deployed on edge devices, Zero-Shot Quantization for OD (ZSQ-OD) aims to quantize these models when access to the original training data is prohibited. Existing research on Zero-Shot Quantization-Aware Training (QAT) for OD synthesizes training sets through noise optimization. However, this approach struggles to maintain performance in low-bit regions. In this paper, we introduce **GoodQ** (**G**enerative **o**ff-the-shelf models for **o**bject **d**etector **Q**uantization), a QAT pipeline that utilizes off-the-shelf generative models to construct a training set. We first identify three challenges that arise when introducing a generative model to the ZSQ-OD task: 1) each image contains dense information with multiple instances, 2) the class-wise distribution in the original dataset is imbalanced, and 3) the pseudo-labels assigned to the generated images can potentially act as noisy signals during QAT. GoodQ addresses these challenges by 1) introducing an Information-Dense Prompting strategy to generate multi-instance images, 2) applying Intrinsic Distribution-Aware Selection to match the pretrained class distribution, and 3) employing Teacher-guided Adaptive Noise Reduction to mitigate noise arising from the QAT process. Our framework achieves state-of-the-art performance in low-bit ZSQ (W4A4) and extends quantization to extreme bit-widths (W3A3). Furthermore, we conduct an extensive analysis to uncover the underlying factors contributing to the efficacy of GoodQ. The constructed dataset is available at 🤗 `GoodQ`.

Keywords: Zero-Shot Quantization, Object Detection, Quantization Aware Training

1 Introduction

Quantization has become a standard practice for deploying deep learning models in resource-constrained environments. Quantization-Aware Training (QAT) effectively compresses full-precision models to lower bit-widths by fine-tuning

† Corresponding author.
* Equal contribution.

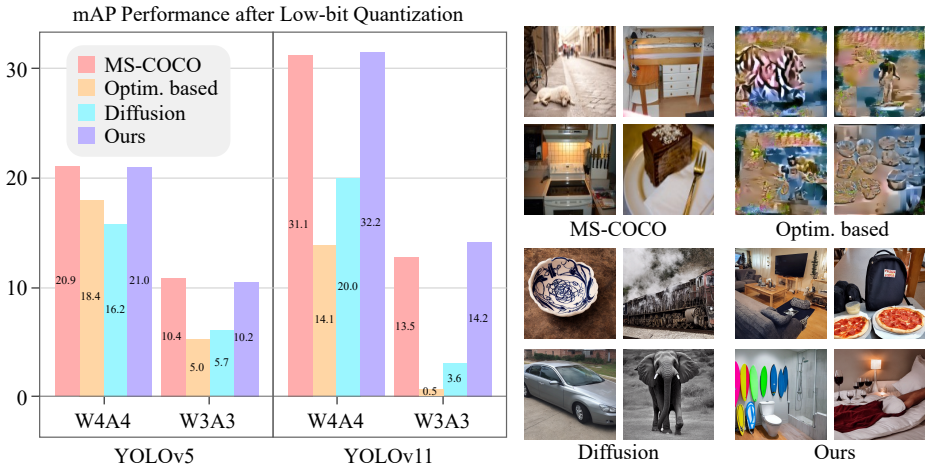

Fig. 1: Left Performance degradation across different quantization bit-widths for the optimization-based method (Optim. based) and our proposed approach, relative to the real dataset baseline. **Right** A visual comparison of images from the real dataset alongside those generated with optimization, diffusion, and our method.

them on a small training set selected from the original training dataset. However, access to the original data is often infeasible in real-world applications due to privacy regulations. To address this, Zero-Shot Quantization (ZSQ) has emerged as a promising paradigm [3–5,11,13,26]. ZSQ synthesizes a set of proxy samples to substitute for the original dataset during training. While early ZSQ methods were based on noise optimization relying on batch normalization (BN) statistic matching [3], subsequent research has shifted toward enhancing the diversity of these synthetic samples to better approximate the distribution of the original dataset [4,13,25]. Notably, recent works such as GenQ [17] and MixQ [20] have demonstrated that integrating off-the-shelf generative models is an effective approach for diversifying synthetic training sets in classification tasks.

As in image classification, we hypothesize that data diversity is crucial for maintaining quantization quality in Object Detection (OD) tasks. To verify this hypothesis, we conducted a preliminary experiment. Using the OD-specific ZSQ algorithm introduced in TSOD [12], we compared three dataset configurations: a subset of the real MS-COCO dataset [18], the noise-optimized synthetic dataset proposed in TSOD, and a diffusion-generated dataset[1]. Figure 1 illustrates these results by showing the mAP scores of the three configurations. As shown in the figure, the noise-optimized synthetic dataset exhibits a limitation: failure in low-bit regions. While the noise-optimized configuration suffers an excessive performance drop, the diverse dataset generated by an off-the-shelf diffusion model achieves relatively better quantization performance at low-bit region. This obser-

[1] Following GenQ, we used the prompt, "A photo of a {D} {C}," where {D} is a CLIP ImageNet template and {C} is a randomly sampled class label.

vation supports our hypothesis that promoting dataset diversity is important in OD, motivating us to leverage diffusion models to enhance quantization quality.

Based on the aforementioned observation, this paper aims to achieve further improvements by addressing three challenges specific to OD: 1) Unlike image classification, OD requires each image to contain dense semantic information, often encompassing multiple object categories. 2) The class-wise distribution in OD datasets is inherently imbalanced, making it difficult to accurately reproduce these distributional characteristics. 3) When conducting QAT with training data synthesized by generative models, the pseudo-labels assigned to these images may introduce noisy supervision signals, potentially misleading the QAT process.

To address these challenges, we propose **GoodQ** (**G**enerative **o**ff-the-shelf models for **o**bject **d**etector **Q**uantization), a pipeline that leverages off-the-shelf generative models tailored specifically for the ZSQ-OD task. GoodQ tackles each challenge through 1) Information-Dense Prompting, 2) Intrinsic Distribution-Aware Selection, and 3) Teacher-guided Adaptive Noise Reduction. These components enable GoodQ to effectively utilize off-the-shelf generative models within the context of ZSQ-OD. To evaluate the effectiveness of the proposed method, we conduct extensive experiments on YOLO-based OD models [8,9], demonstrating that GoodQ excels in low-bit regions while maintaining competitive performance at higher bit-widths. Beyond quantitative results, we also provide comprehensive analyses to explain the effectiveness of GoodQ.

In summary, our main contributions are as follows:

- We investigate the utilization of off-the-shelf generative models for the ZSQ-OD task and identify the key challenges associated with this approach.
- We propose GoodQ, a novel framework designed to tailor generative models specifically for OD tasks, effectively mitigating the identified challenges.
- Through extensive experiments and analysis, we demonstrate the underlying factors contributing to the effectiveness of GoodQ.

2 Related Works

2.1 Quantization

Quantization reduces the memory footprint and enables efficient inference by mapping high-precision tensors to low-bit integers. Quantization-Aware Training (QAT) [2, 6, 10, 14] utilizes a small training set to determine quantization parameters and update model weights. Specifically, the weights are first mapped to integers using initial quantization parameters and then fine-tuned on the curated training set. The simulated quantization of a high-precision tensor x to a b-bit representation is formulated as follows:

$$\hat{x} = s \times \left\lfloor \mathrm{clip}\Big(\frac{x}{s},\, n,\, p\Big) \right\rceil, \tag{1}$$

where $n = -2^{b-1}$ and $p = 2^{b-1} - 1$ for symmetric weight quantization, while $n = 0$ and $p = 2^b - 1$ for asymmetric activation quantization. Here, s is a

learnable scaling parameter, $\lfloor \cdot \rceil$ denotes the rounding operation, and clip$(\cdot)$ is a function that clamps values to the range $[n, p]$.

2.2 Zero-Shot Quantization

While the majority of QAT methods curate their training sets from the original dataset, there are cases where access to these data is restricted due to privacy regulations and security concerns. To address this limitation, researchers have investigated quantization methods that employ synthetic datasets. This line of research is referred to as Zero-Shot Quantization (ZSQ) [3, 11, 17, 20, 22].

Synthetic training dataset for ZSQ is primarily generated in three ways: 1) through noise optimization, 2) by training a GAN-based generator, and 3) by leveraging an off-the-shelf (mostly diffusion-based) generative model. Noise optimization methods utilize the pretrained model to directly optimize learnable inputs initialized with Gaussian noise [3, 11]. GAN-based methods, on the other hand, train an adversarial network so that the generator can accurately approximate the internal knowledge of the model [1, 24, 25]. However, both noise optimization and GAN-based methods struggle to maintain performance when quantized to lower bit-widths due to a lack of data diversity. To tackle this issue, methods that leverage off-the-shelf generative models have emerged [15, 17, 20], achieving state-of-the-art performance.

2.3 Zero-shot Quantization for Object Detection

However, most previous research on ZSQ has primarily focused on image classification, largely overlooking other tasks such as Object Detection (OD)—a task in high demand for deployment on edge devices in real-world scenarios. TSOD [12] is a pioneering work that addresses ZSQ specifically for OD (ZSQ-OD) by incorporating a task-specific QAT objective, $\mathcal{L}_{\text{detect}}$, alongside the task-agnostic term, $\mathcal{L}_{\text{distill}}$. It proposes a noise-optimization-based method to synthesize a training set by leveraging task-specific information, achieving superior performance. Despite this, TSOD exhibits a critical limitation: it struggles to maintain robustness in low-bit quantization regions (e.g., W4A4). This performance degradation at low bit-widths hinders its applicability when a quantized model needs to be deployed in resource-constrained environments.

2.4 Position of Our Work

As empirically demonstrated in Figure 1, the performance degradation of TSOD in low-bit regions can potentially be mitigated by utilizing synthetic training sets generated by off-the-shelf models. Building on this observation, we propose a ZSQ-OD pipeline that leverages such generative models in a manner tailored specifically for OD tasks. To achieve this, we first identify three key challenges that must be addressed. By systematically resolving these challenges, we formulate our comprehensive ZSQ-OD pipeline: GoodQ.

3 Three Challenges of ZSQ-OD

In this section, we introduce three challenges that must be addressed when adapting off-the-shelf generative models specifically for the OD task: two arising during the training set generation stage, and one during the Quantization-Aware Training (QAT) stage. These challenges can be summarized as follows:

1. **Information Density Challenge**: A single image may contain multiple instances, and thus multiple labels.
2. **Class-wise Imbalance Challenge**: Each class may contain a different number of instances due to the multi-label nature of OD datasets.
3. **Pseudo-label Challenge**: The pseudo-labels assigned to each image can introduce noisy supervision signals during the QAT process.

3.1 Information Density Challenge

A primary characteristic of OD datasets is the high density of information present within each image. For instance, the MS-COCO 2017 dataset [18]—one of the most widely used OD datasets—contains an average of 7.7 bounding boxes (BBoxes) per image. While object sparsity has not been a critical issue for single-label tasks such as image classification, OD is a multi-label task that inherently depends on both categorical information and spatial localization (e.g., BBoxes). Consequently, there is a compelling need to generate images with information-rich content to adequately satisfy the requirements of OD tasks.

3.2 Class-wise Imbalance Challenge

Another challenge arises from the extreme class-wise distribution skew inherent in OD datasets. For instance, in the full MS-COCO dataset, the proportions of bounding boxes for the most frequent (head) and least frequent (tail) classes are roughly 30.52% and 0.02%, respectively. This represents a far more severe long-tail imbalance than is typically observed in image classification tasks. Consequently, a naive application of generative models for image generation is likely to overlook or misrepresent this heavily skewed distribution.

3.3 Pseudo-Label Challenge

The final challenge arises from the pseudo-labels assigned to the generated images. The objective of QAT is to guide the initially quantized model to match the behavior of the full-precision model using a small training set. The detection-specific term of the ZSQ-OD objective ($\mathcal{L}_{\text{detect}}$) takes two inputs: $\mathbf{y}$, which is the target serving as guidance, and $\hat{\mathbf{y}}$, which is the output of the quantized model being optimized. The objective is formulated as follows:

$$\mathcal{L}_{\text{detect}}(\hat{\mathbf{y}}, \mathbf{y}) = \lambda_{\text{b}} \underbrace{(1 - \text{CIoU}(\hat{\mathbf{y}}_{\text{b}}, \mathbf{y}_{\text{b}}))}_{\mathcal{L}_{\text{box}}} + \lambda_{\text{o}} \underbrace{\text{BCE}(\hat{y}_{\text{o}}, y_{\text{o}})}_{\mathcal{L}_{\text{obj}}} + \lambda_{\text{c}} \underbrace{\text{BCE}(\hat{\mathbf{y}}_{\text{c}}, \mathbf{y}_{\text{c}})}_{\mathcal{L}_{\text{cls}}}. \quad (2)$$

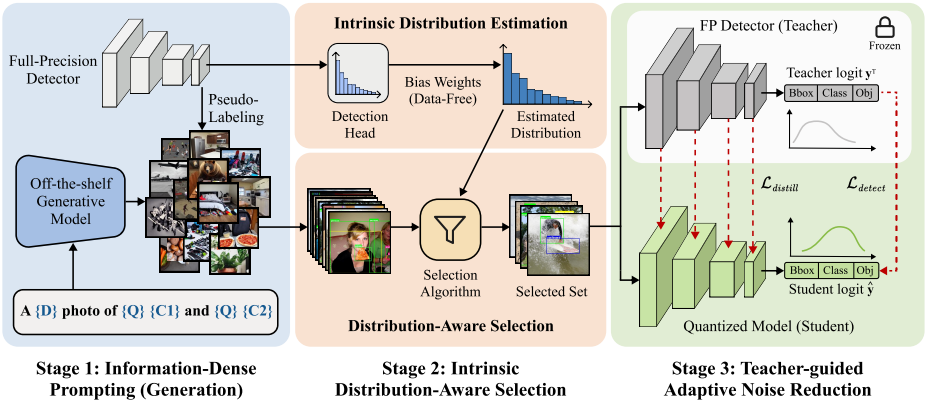

Fig. 2: Overview of the GoodQ pipeline. **Stage 1** Information-Dense Prompting constructs an image pool tailored for Object Detection using an off-the-shelf generative model. **Stage 2** Intrinsic Distribution-Aware Selection estimates the inherent class distribution and selects a representative training dataset from the image pool. **Stage 3** Teacher-guided Adaptive Noise Reduction utilizes teacher-generated soft predictions as targets to mitigate label noise during the QAT process.

TSOD employs $\mathbf{y}^{\mathrm{GT}} \triangleq (\mathbf{y}_b^{\mathrm{GT}}, y_o^{\mathrm{GT}}, \mathbf{y}_c^{\mathrm{GT}})$ as $\mathbf{y}$, where $\mathbf{y}_b^{\mathrm{GT}} \in \mathbb{R}^4$, $y_o^{\mathrm{GT}} \in \mathbb{R}$, and $\mathbf{y}_c^{\mathrm{GT}} \in \mathbb{R}^{|\mathcal{C}|}$ represent the ground truths for bounding box coordinates, objectness score, and class probabilities, respectively. However, when utilizing an off-the-shelf generative model, the generated images are unlabeled, requiring a pseudo-labeling process to substitute for $\mathbf{y}^{\mathrm{GT}}$. Because the pseudo-labeling process generally enforces one-hot style hard labels, it introduces potential noise that inherently depends on the performance of the model utilized for labeling.

4 Method

To address the aforementioned challenges, we propose GoodQ, a ZSQ-OD pipeline that leverages off-the-shelf generative models. Our method consists of two primary steps: 1) constructing a training dataset tailored specifically for QAT in OD, and 2) performing the QAT process using the constructed dataset. Figure 2 illustrates the overview of our proposed method. In the following sections, we provide a detailed explanation of our approach across three stages.

4.1 Stage 1. Information-Dense Prompting

Unlike image classification, ZSQ-OD inherently necessitates information-dense images where multiple object categories and BBoxes coexist within a single scene. To address the Information Density Challenge, we propose Information-Dense Prompting (IDP) to better reflect the characteristics of real-world OD datasets. IDP employs the following template:

Algorithm 1 Distribution-Aware Selection

Require: Image pool $\mathcal{I}$, target size K, target ratios $\{\hat{r}_c\}$, BBox counts $B_{i,c}$, $c \in \mathcal{C}$
Ensure: Selected image set $\mathcal{S}$

$\mathcal{S} \leftarrow \emptyset$
while $|\mathcal{S}| < K$ and $\mathcal{I} \neq \emptyset$ **do**
  #**Step 1: Target Distribution Estimation & Candidate Update**
  Compute current ratios $\{p_c\}$ from current $\mathcal{S}$
  for $c \in \mathcal{C}$ **do**
    Compute gaps $g_c \leftarrow \hat{r}_c - p_c$ ▷ $g_c > 0$ for under-selected class c
    $\mathcal{A}_c \leftarrow \{i \in \mathcal{I} \mid B_{i,c} > 0\}$ ▷ Set of every image with class c
  end for
  #**Step 2: Priority-Based Target Class Selection**
  $\mathcal{C}_{\text{select}} \leftarrow \{c \in \mathcal{C} \mid g_c > 0 \text{ and } \mathcal{A}_c \neq \emptyset\}$ ▷ Get under-selected class
  if $\mathcal{C}_{\text{select}} \neq \emptyset$ **then**
    $c^* \leftarrow \arg\min_{c \in \mathcal{C}_{\text{select}}} |\mathcal{A}_c|$ ▷ least remaining class
  else
    $c^* \leftarrow \arg\max_{c:\, \mathcal{A}_c \neq \emptyset} g_c$ ▷ least over-selected class
  end if
  #**Step 3: BBox-Driven Image Selection & Update**
  $i^* \leftarrow \arg\max_{i \in \mathcal{A}_{c^*}} \sum_{c \in \mathcal{C}} B_{i,c}$ ▷ Choose i^* with the most BBox
  $\mathcal{S} \leftarrow \mathcal{S} \cup \{i^*\}, \quad \mathcal{I} \leftarrow \mathcal{I} \setminus \{i^*\}$ ▷ Move i^* to selected image set
end while
return $\mathcal{S}$ (pad randomly from $\mathcal{I}$ if $|\mathcal{S}| < K$)

$$\text{Prompt} = \text{“A \{D\} photo of \{Q\} \{C1\} and \{Q\} \{C2\}”}, \tag{3}$$

where $\{C_1\}$ and $\{C_2\}$ denote class labels, $\{D\}$ represents a descriptive adjective (e.g., clean, nice), and $\{Q\}$ is a quantitative adjective (e.g., multiple, many). A detailed list of the prompts used in our experiments is provided in the Appendix. By leveraging IDP, we generate a pool of 160k candidate images with a high density of BBoxes tailored for ZSQ-OD. Subsequently, we employed a pre-trained detection model to generate bounding box annotations, thereby assigning pseudo-labels to the synthesized images.

4.2 Stage 2. Intrinsic Distribution-Aware Selection

OD datasets exhibit a more inherently skewed class distribution compared to image classification. To address the Class-wise Imbalance Challenge, we propose Intrinsic Distribution-Aware Selection (IDAS). IDAS aligns the distribution of the generated images with that of the original dataset through two steps: 1) estimating the distribution of the real dataset in a data-free setting, where the original training images and labels are entirely inaccessible, and 2) selecting images from the generated pool to match this estimated class-wise distribution.

Intrinsic Distribution Estimation In data-free scenarios, access to the class distribution of the original dataset is strictly restricted. Consequently, the distribution must be estimated from the only available source: the pre-trained model weights. The bias terms of the detector may implicitly capture information about the class prior. Therefore, to reliably estimate the target distribution from this limited information source, we leverage the bias parameters of the detection head. Specifically, the target class-wise ratio is computed as follows:

$$\hat{w}_c = \bar{b_c} - \min_{i \in \mathcal{C}} \bar{b_i}, \quad \hat{r}_c = \frac{\max\big(\hat{w}_c / \sum_{k \in \mathcal{C}} \hat{w}_k,\ \epsilon\big)}{\sum_{j \in \mathcal{C}} \max\big(\hat{w}_j / \sum_{k \in \mathcal{C}} \hat{w}_k,\ \epsilon\big)}, \tag{4}$$

where $\bar{b}_c$ denotes the average bias value for class c extracted from the pre-trained detection head, $\mathcal{C}$ is the set of all object classes, and ϵ is a small constant serving as a correction factor to prevent zero target ratios during normalization. By applying Equation (4), we derive the set of target ratios $\mathcal{R} \triangleq \{\hat{r}_c \mid c \in \mathcal{C}\}$. Notably, the distribution estimated via our approach exhibits a strong Pearson correlation ($r = 0.87$) with the ground-truth class distribution of the MS-COCO dataset. This high correlation provides empirical validation for our proposed estimation strategy and demonstrates its effectiveness in accurately capturing the target distribution even in a data-free setting.

Distribution-Aware Selection Leveraging the pre-computed target ratios $\mathcal{R}$ from the previous section, we introduce Distribution-Aware Selection (DAS) to curate the generated image pool to better reflect the estimated target distribution. Given the pool of images generated in Section 4.1, we perform a greedy selection process until the set of selected images $\mathcal{S}$ reaches the target size K.The selection process follows these steps in each iteration: we calculate the class distribution p_c of the images in $\mathcal{S}$ and compute the discrepancy between p_c and $\hat{r}_c$ (Step 1). We then rank the classes by prioritizing those with the largest discrepancy, while incorporating an additional preference for those with fewer remaining candidate images (Step 2). Based on this ranking, a target class is chosen. Among the images that contain the chosen class, we select the one with the largest number of detected BBoxes and update $\mathcal{S}$ (Step 3). The complete procedure is detailed in Algorithm 1.

4.3 Stage 3. Teacher-guided Adaptive Noise Reduction

Assigning pseudo-labels to diffusion-generated images using a pre-trained detection model inevitably introduces label noise. Therefore, it is inappropriate to directly apply the conventional Object Detection QAT loss $\mathcal{L}_{\text{detect}}$ in Equation (2) to our setting. To mitigate the adverse effects of such noisy supervision, Teacher-guided Adaptive Noise Reduction (TANR) bypasses pseudo-labels entirely by constructing $\mathcal{L}_{\text{detect}}$ directly from the teacher's soft predictions as targets. Specifically, we replace the pseudo-label with the soft label $\mathbf{y}^{\mathcal{T}} \triangleq (\mathbf{y}_b^{\mathcal{T}}, y_o^{\mathcal{T}}, \mathbf{y}_c^{\mathcal{T}})$ produced by the full-precision teacher model. Furthermore, we adopt QFocal-based Adaptive Weighting (AW) [16], which yields the revised loss:

Table 1: Quantization Aware Training (QAT) results on YOLOv5/YOLOv11 using MS-COCO validation set. We fix the size of the calibration set for all methods to 2k and compare the results on mAP / mAP50 score.

	Real Data			mAP / mAP50					
Method	Image	Label	Prec.	YOLOv5-s	YOLOv5-m	YOLOv5-l	YOLOv11-s	YOLOv11-m	YOLOv11-l
Pre-trained	✓	✓	FP	37.4 / 56.8	45.4 / 64.1	49.0 / 67.3	47.0 / 65.0	51.5 / 70.0	53.4 / 72.5
LSQ	✓	✓		32.6 / 51.2	38.4 / 57.1	40.0 / 58.4	44.0 / 60.8	47.6 / 64.5	48.8 / 65.8
LSQ+	✓	✓		32.2 / 51.0	38.0 / 56.9	39.4 / 58.5	43.8 / 60.7	47.8 / 64.7	48.5 / 65.3
TSOD	×	✓	W8A8	35.4 / 54.0	**42.9** / **61.4**	**45.7** / **63.8**	**45.7** / **61.7**	**50.1** / 65.6	51.5 / 66.9
TSOD†	×	×		**35.9** / **54.7**	37.1 / 53.7	30.4 / 46.0	43.8 / 58.2	44.7 / 56.9	44.6 / 55.7
GoodQ	×	×		35.0 / 53.6	42.1 / 60.1	45.6 / 63.4	45.3 / **61.7**	50.0 / **66.3**	**51.8** / **67.8**
LSQ	✓	✓		29.8 / 47.9	36.2 / 54.3	38.6 / 56.7	41.5 / 58.3	45.0 / 61.9	45.8 / 62.5
LSQ+	✓	✓		29.0 / 47.6	35.6 / 53.8	38.0 / 57.2	41.6 / 58.2	44.8 / 61.7	45.9 / 62.8
TSOD	×	✓	W6A6	32.3 / **50.6**	**40.4** / **58.4**	43.5 / **61.4**	42.4 / 57.9	45.6 / 60.7	46.4 / 60.6
TSOD†	×	×		32.0 / 50.0	32.7 / 48.1	28.8 / 43.1	38.1 / 49.9	34.5 / 44.2	34.1 / 41.7
GoodQ	×	×		**32.5** / **50.6**	40.0 / 58.0	**43.6** / **61.4**	**43.3** / **59.4**	**47.7** / **63.9**	**49.3** / **65.0**
LSQ	✓	✓		19.2 / 34.4	27.3 / 44.1	31.0 / 48.2	29.2 / 43.7	35.3 / 50.4	36.3 / 51.9
LSQ+	✓	✓		19.0 / 34.0	27.4 / 44.5	31.1 / 48.2	29.3 / 43.4	34.8 / 50.1	36.3 / 51.9
TSOD	×	✓	W4A4	18.4 / 32.3	28.7 / 45.0	33.5 / 50.3	14.1 / 21.4	16.7 / 23.7	20.1 / 28.6
TSOD†	×	×		15.0 / 27.9	20.3 / 32.0	19.3 / 30.8	6.58 / 9.61	5.95 / 9.00	6.22 / 8.89
GoodQ	×	×		**21.0** / **36.1**	**30.7** / **47.6**	**34.8** / **52.0**	**32.2** / **45.9**	**37.4** / **51.6**	**39.4** / **53.6**
LSQ	✓	✓		8.48 / 17.7	17.6 / 31.1	20.9 / 35.4	11.7 / 19.6	19.9 / 31.1	19.2 / 29.7
LSQ+	✓	✓		9.44 / 19.6	17.9 / 31.3	21.5 / 35.9	11.6 / 19.1	18.3 / 28.6	**20.4** / **31.6**
TSOD	×	✓	W3A3	5.0 / 9.9	15.0 / 26.5	21.1 / 33.8	0.5 / 0.7	0.8 / 1.5	0.6 / 1.1
TSOD†	×	×		2.56 / 5.20	6.41 / 11.9	6.84 / 12.4	0.18 / 0.40	0.50 / 0.98	0.22 / 0.43
GoodQ	×	×		**10.2** / **19.7**	**20.2** / **33.0**	**25.4** / **39.1**	**14.2** / **22.1**	**21.6** / **31.8**	19.5 / 28.7

$$\mathcal{L}_{\text{detect}}(\hat{\mathbf{y}}, \mathbf{y}^{\mathcal{T}}) = \lambda_{\text{b}}\mathcal{L}_{\text{box}} + |y_{\text{o}}^{\mathcal{T}} - \hat{y}_{\text{o}}|^{\gamma}\lambda_{\text{o}}\mathcal{L}_{\text{obj}} + \sum_{\text{c}_i \in \mathcal{C}} |y_{\text{c}_i}^{\mathcal{T}} - \hat{y}_{\text{c}_i}|^{\gamma}\lambda_{\text{c}}\mathcal{L}_{\text{cls}}, \tag{5}$$

where γ denotes a scaling factor. $\mathbf{y}_c^{\mathcal{T}}$ and $\hat{\mathbf{y}}_c$ denote the teacher's and student's class-probability vectors over the class set $\mathcal{C}$, respectively. For each class $c_i \in \mathcal{C}$, $y_{c_i}^{\mathcal{T}}$ and $\hat{y}_{c_i}$ denote the i-th elements of $\mathbf{y}_c^{\mathcal{T}}$ and $\hat{\mathbf{y}}_c$, corresponding to the predicted probability of class c_i. AW guides the QAT process by enabling the student model to focus on instances that are more susceptible to quantization-induced noise. Finally, we combine our $\mathcal{L}_{\text{detect}}$ with the conventional $\mathcal{L}_{\text{distill}}$ to formulate the overall QAT objective.

5 Experiments

5.1 Experimental Setting

While GoodQ is applicable regardless of the model architecture, we mainly focus on the YOLO family [23], a single-stage detection architecture widely deployed on edge devices. More specifically, we conducted experiments using YOLOv5 [8]

Table 2: Ablation studies on (a) training dataset construction process and (b) quantization-aware training loss design on mAP / mAP50 score.

(a) Dataset Construction

Generation	W8A8	W4A4
Diffusion	33.5 / 52.1	16.0 / 28.8
+IDP	34.3 / 53.5	19.3 / 33.8
+IDAS	34.4 / 52.9	19.8 / 34.4
+IDP + IDAS	**35.0** / **53.6**	**21.0** / **36.1**

(b) QAT loss

Method	W8A8	W4A4
Base	33.9 / 52.8	20.6 / 35.8
+TANR	**35.0** / **53.6**	**21.0** / **36.1**

and YOLOv11 [9] models pre-trained on the MS-COCO 2017 dataset [18], following TSOD [12]. We adopted the LSQ [6] scheme to quantize both weights and activations. To demonstrate the effectiveness of our method, we compared our results with LSQ [6], LSQ+ [2], and TSOD [12]. Both LSQ and LSQ+ results are reported on a subset of the MS-COCO dataset, whereas the TSOD results are reported on a noise-optimized dataset synthesized using real labels. All experiments were conducted using a training set of size 2k. For a fair comparison, no data augmentation was applied.

5.2 Results

Table 1 presents the evaluation results of YOLOv5 and YOLOv11 on the COCO dataset across various QAT and ZSQ-OD methods. The best-performing results are in **bold**, and the second-best performance is underlined. We additionally report a label-free variant of TSOD, which we denote as TSOD† As shown in the table, GoodQ achieves performance comparable to existing approaches in high-bit regimes (e.g., W8A8 and W6A6). While competing methods suffer severe degradation in extreme low-bit regimes (e.g., W4A4 and W3A3), GoodQ remains substantially more robust, exhibiting only a minimal performance drop. This highlights the advantage of leveraging off-the-shelf generative models for ZSQ-OD over traditional optimization-based approaches such as TSOD. Notably, even without access to the original data, our zero-shot approach markedly outperforms TSOD in these low-bit settings, despite TSOD relying on ground-truth labels during QAT optimization. Overall, these results validate the effectiveness of our framework for ZSQ-OD without requiring any access to the original dataset.

5.3 Ablation Studies

We conducted three ablation studies to examine the effect of each component of GoodQ independently. Specifically, we varied 1) the training dataset construction process, 2) the QAT loss design, and 3) the size of the image pool. All experiments were conducted with the YOLOv5-s model.

Table 3: Ablation study on the size of the image pool across different quantization bit-widths compared on mAP / mAP50 score.

# Image Pool	Bit-width			
	W8A8	W6A6	W4A4	W3A3
2k	34.3 / 53.5	31.8 / 50.1	19.3 / 33.8	7.76 / 15.2
16k	34.9 / 53.6	32.4 / 50.7	20.2 / 35.3	9.24 / 17.2
32k	**35.2** / **53.9**	32.5 / **50.9**	21.1 / **36.7**	9.41 / 17.5
64k	35.1 / 53.8	**32.6** / **50.9**	**21.3** / 36.5	9.20 / 17.0
160k (Ours)	35.0 / 53.6	32.5 / 50.6	21.0 / 36.1	**10.20** / **19.7**

Ablation on Constructing Training Dataset Table 2-(a) presents an ablation study on the training-dataset construction process for QAT under W8A8 and W4A4. DIFFUSION in the table denotes the baseline, where images generated using GenQ prompts originally designed for classification [17] are used as the training set[2]. As shown in the table, both Information-Dense Prompting (IDP) and Intrinsic Distribution-Aware Selection (IDAS) improve ZSQ performance across all bit-widths. Combining the two provides a further boost. Overall, these results confirm that both IDP and IDAS are effective components that reflect the intrinsic characteristics of Object Detection datasets.

Ablation on QAT Loss Design Table 2-(b) presents an ablation study on the loss functions used for QAT. BASE in the table uses the losses proposed in TSOD, specifically combining $\mathcal{L}_{\text{distill}}$ with the task-specific loss $\mathcal{L}_{\text{detect}}$ computed using one-hot labels. As shown in the table, replacing $\mathcal{L}_{\text{detect}}$ with Teacher-guided Adaptive Noise Reduction (TANR) improves performance across all bit regimes. These results indicate that TANR consistently enhances QAT performance.

Ablation on the Number of Image Pool In this paper, we use an image pool of size 160k generated by a generative model. In this section, we conduct an ablation study to analyze how the image-pool size affects ZSQ-OD performance under different quantization bit settings. Table 3 summarizes the results for various image-pool sizes. As shown in the table, increasing the image-pool size generally improves performance, highlighting the benefit of a richer and more diverse set of synthetic images. Notably, the performance gains are more pronounced in low-bit settings, where data diversity plays a more critical role in preserving accuracy under quantization.

[2] We use the GenQ prompt, "A photo of a {D} {C}," where {D} is a CLIP ImageNet template and {C} is a randomly sampled class label.

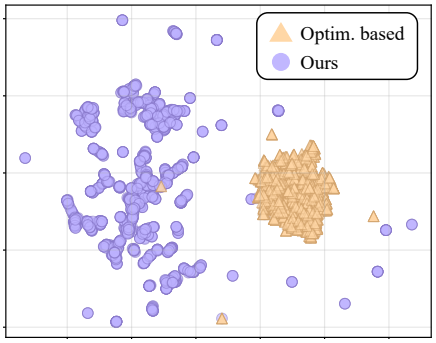

Fig. 3: UMAP visualization of CLIP embeddings for the optimization-based synthesized dataset and Ours.

Table 4: Cosine distance measured on the optimization-based synthesized training dataset and ours across various quantization settings. Higher values indicate greater diversity. Ours show higher diversity across all bit rates.

Dataset	FP	W8A8	W4A4
Optim.	0.539	0.535	0.133
Ours	**0.665**	**0.665**	**0.248**

6 Analysis

In this section, we provide analyses to answer why GoodQ is effective. We first analyze GoodQ from the perspective of dataset generation and then examine it in the context of the QAT process. All analyses are performed using the YOLOv5-s.

6.1 Analysis on Dataset Generation Perspective

We first analyze the dataset generation in GoodQ by examining two aspects: 1) the importance of data diversity in the low-bit regime and 2) the need to align the class distribution of the constructed training dataset. For the first analysis, we compare the noise-optimization-based synthetic dataset with the dataset constructed using a generative model. For the second analysis, to evaluate the importance of distribution alignment, we incorporate different image selection methods and measure QAT performance on the resulting training sets.

Finding 1. Diversity Matters in Low-bit As demonstrated in Table 1, performing ZSQ-OD with a diffusion-generated training dataset yields superior performance in extreme low-bit regimes (e.g., W4A4 and W3A3) compared to noise-optimization-based methods such as TSOD. To understand the source of this advantage, we analyze the feature representations of the two datasets. Figure 3 shows a UMAP [19] visualization of their CLIP [21] embeddings, indicating that the diffusion-generated dataset is substantially more dispersed. This observation is quantitatively supported by Table 4, which reports the average pairwise cosine distance[3] computed from embeddings extracted by both full-precision and

[3] Let $(x_i)_{i=1}^N$ denote embeddings extracted from N images. We compute the cosine distance for each image pair (i, j) as $d_{\cos}(x_i, x_j) = 1 - \cos(x_i, x_j)$. The final number reported in the table was calculated by averaging over every pairwise cosine distance $d_{\cos}(x_i, x_j)$ (typically $i < j$).

Table 5: Comparison of the class-wise BBox distribution between the real COCO training dataset and the distribution resulting from diverse selection methods.

Selection	W8A8	W4A4	#BBox	KL	L1
Random	34.3 / 53.5	19.3 / 33.8	7917	0.48	0.77
Many BBox	33.5 / 51.7	17.5 / 30.7	**72786**	1.25	1.14
IDAS	**35.0**/ **53.6**	**21.0** / **36.1**	25333	**0.2616**	**0.6116**

quantized models. These results confirm the superior diversity of the diffusion-generated images. Such enhanced diversity appears particularly beneficial under severe quantization constraints. By covering a broader region of the feature space, the synthetic dataset can help the heavily quantized student model capture the knowledge of the full-precision teacher more effectively [7]. Overall, our analysis suggests that maximizing training-data diversity is an important factor for robust low-bit ZSQ-OD performance.

Finding 2. Class Distribution Matters As illustrated in Table 2-(a), estimating and curating a dataset that matches the original skewed class-wise distribution improves QAT performance in the low-bit regime. To further investigate the importance of aligning generated data with the class-wise bounding-box distribution, we apply three selection algorithms after constructing an image pool using Information-Dense Prompting (IDP). Table 5 provides a comprehensive comparison of performance and statistics for random selection (Random), naive selection based on the number of bounding boxes (Many BBox), and IDAS. The superior performance of IDAS over Random suggests that accurate distribution matching directly improves QAT performance. Notably, the comparison with Many BBox indicates that the gains are not merely a byproduct of selecting images with more bounding boxes. As shown in the table, this naive strategy degrades performance.

These results suggest that explicitly reflecting the intrinsic characteristics of the original dataset during training-data construction is crucial for successful ZSQ-OD. As shown in the same table, the dataset selected by IDAS matches the original distribution more closely than Random, as evidenced by lower KL divergence and L1 distance relative to the original class-wise distribution[4]. In contrast, Many BBox selection deviates further from the original distribution. Overall, these findings confirm that the QAT training dataset should reflect the true class-wise distribution.

6.2 Analysis on QAT Perspective

Next, we analyze the QAT process in GoodQ by examining the importance of mitigating pseudo-label noise before QAT. Specifically, we compare the impact

[4] The class-wise BBox distribution of the original OD dataset is computed as the ratio of BBoxes belonging to each class to the total number of BBoxes in the training set.

Table 6: Comparison of the quantization performance before and after applying TANR to both ours dataset and optimization-based images. Ours indicates the resulting set gathered by applying IDAS to the diffusion image pool.The results in the table report performance in terms of mAP/mAP50.

Precision	Method	Ours	Optim.
FP	-	37.4 / 56.8	
W8A8	Base	33.9 / 52.8	35.4 / 54.0
	+TANR	35.0 / 53.6	35.7 / 54.4
W4A4	Base	20.6 / 35.8	18.4 / 32.3
	+TANR	21.0 / 36.1	18.3 / 32.1

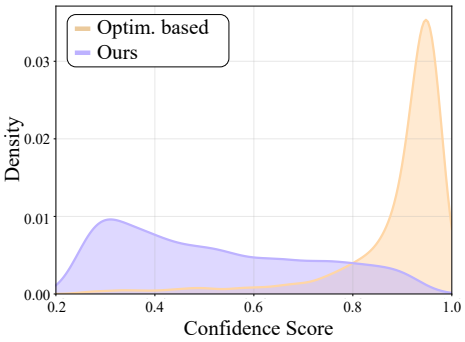

Fig. 4: Classification confidence distributions of localized objects under different generation configurations are shown. For improved visualization, the overall distributions are smoothed.

of Teacher-guided Adaptive Noise Reduction (TANR) across two distinct training datasets: one generated by a diffusion model and the other constructed via optimization-based synthesis.

Finding 3. Teacher Noise Matters Table 6 presents a performance comparison when TANR is applied to training datasets constructed via optimization-based synthesis versus our diffusion-based generation. As shown in the table, TANR consistently improves performance on the diffusion-generated dataset, whereas its gains on the optimization-based baseline are marginal. The reason for this discrepancy is illustrated in Figure 4, showing that diffusion-generated images yield lower confidence scores[5] than images synthesized by optimization-based methods. These results indicate that TANR is effective for addressing the challenges that arise when using generative models for ZSQ-OD.

7 Conclusion

In this paper, we introduce GoodQ (**G**enerative **o**ff-the-shelf models for **o**bject **d**etector **Q**uantization), a novel framework designed to improve Zero-Shot Quantization for Object Detection (ZSQ-OD) by leveraging off-the-shelf generative models in a task-specific manner. For Object Detection, GoodQ accounts for challenges such as the need for high-density information within a single frame and highly skewed class-wise distributions. Moreover, generative models introduce additional complexity because pseudo-labels are inherently noisy. To address these challenges, GoodQ incorporates three synergistic components: Information-Dense Prompting (IDP), Intrinsic Distribution-Aware Selection (IDAS), and

[5] The confidence score is computed as the product of the objectness score and the class confidence: $p_{\text{conf}} = \sigma(\hat{y}_o) \cdot \max(\sigma(\hat{\mathbf{y}}_c))$.

Teacher-guided Adaptive Noise Reduction (TANR). Extensive experiments across detection models show that GoodQ substantially mitigates performance degradation in challenging low-bit regimes—where optimization-based synthesis methods struggle—while maintaining competitive performance in high-bit regimes. Finally, we provide in-depth analyses across multiple dimensions, offering practical insights for successful ZSQ-OD.

Acknowledgements

This work was supported by the Korean Government through the grants from IITP (RS-2021-II211343, RS-2025-25442338, 26-InnoCORE-01).

References

1. Bai, J., Yang, Y., Chu, H., Wang, H., Liu, Z., Chen, R., He, X., Mu, L., Cai, C., Hu, H.: Robustness-guided image synthesis for data-free quantization. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 10971–10979 (2024)
2. Bhalgat, Y., Lee, J., Nagel, M., Blankevoort, T., Kwak, N.: Lsq+: Improving low-bit quantization through learnable offsets and better initialization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2020)
3. Cai, Y., Yao, Z., Dong, Z., Gholami, A., Mahoney, M.W., Keutzer, K.: Zeroq: A novel zero shot quantization framework. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13169–13178 (2020)
4. Choi, K., Hong, D., Park, N., Kim, Y., Lee, J.: Qimera: Data-free quantization with synthetic boundary supporting samples. Advances in Neural Information Processing Systems **34**, 14835–14847 (2021)
5. Choi, K., Lee, H., Kwon, D., Park, S., Kim, K., Park, N., Choi, J., Lee, J.: Mimiq: Low-bit data-free quantization of vision transformers with encouraging inter-head attention similarity. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 16037–16045 (2025)
6. Esser, S.K., McKinstry, J.L., Bablani, D., Appuswamy, R., Modha, D.S.: Learned step size quantization. arXiv preprint arXiv:1902.08153 (2019)
7. Frank, L., Davis, J.: What makes a good dataset for knowledge distillation? In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23755–23764 (2025)
8. Khanam, R., Hussain, M.: What is yolov5: A deep look into the internal features of the popular object detector. arXiv preprint arXiv:2407.20892 (2024)
9. Khanam, R., Hussain, M.: Yolov11: An overview of the key architectural enhancements. arXiv preprint arXiv:2410.17725 (2024)
10. Kim, J., Bhalgat, Y., Lee, J., Patel, C., Kwak, N.: Qkd: Quantization-aware knowledge distillation. arXiv preprint arXiv:1911.12491 (2019)
11. Kim, M., Kim, J., Kang, U.: Synq: Accurate zero-shot quantization by synthesis-aware fine-tuning. In: The Thirteenth International Conference on Learning Representations (2025)
12. Li, C., Chen, X., Wang, J., Zhao, K., Chen, J.: Task-specific zero-shot quantization-aware training for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22868–22878 (2025)

13. Li, H., Wu, X., Lv, F., Liao, D., Li, T.H., Zhang, Y., Han, B., Tan, M.: Hard sample matters a lot in zero-shot quantization. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 24417–24426 (2023)
14. Li, R., Wang, Y., Liang, F., Qin, H., Yan, J., Fan, R.: Fully quantized network for object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2810–2819 (2019)
15. Li, S., Karmann, M., Urfalioglu, O.: Joint post-training quantization of vision transformers with learned prompt-guided data generation. arXiv preprint arXiv:2602.18861 (2026)
16. Li, X., Wang, W., Wu, L., Chen, S., Hu, X., Li, J., Tang, J., Yang, J.: Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. Advances in neural information processing systems **33**, 21002–21012 (2020)
17. Li, Y., Kim, Y., Lee, D., Kundu, S., Panda, P.: Genq: Quantization in low data regimes with generative synthetic data. In: European Conference on Computer Vision. pp. 216–235. Springer (2024)
18. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
19. McInnes, L., Healy, J., Melville, J.: Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2018)
20. Park, J., Lee, C., Choi, Y., Park, S., Hong, D., Choi, J.: Enhancing generalization in data-free quantization via mixup-class prompting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 4002–4011 (October 2025)
21. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
22. Ramachandran, A., Kundu, S., Krishna, T.: Clamp-vit: Contrastive data-free learning for adaptive post-training quantization of vits. In: European Conference on Computer Vision. pp. 307–325. Springer (2024)
23. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)
24. Xu, S., Li, H., Zhuang, B., Liu, J., Cao, J., Liang, C., Tan, M.: Generative low-bitwidth data free quantization. In: European conference on computer vision. pp. 1–17. Springer (2020)
25. Zhao, K., Zhuang, Z., Zhang, M., Guo, C., Shu, Y., Yang, B.: Enhancing diversity for data-free quantization. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 20969–20978 (2025)
26. Zhong, Y., Lin, M., Nan, G., Liu, J., Zhang, B., Tian, Y., Ji, R.: Intraq: Learning synthetic images with intra-class heterogeneity for zero-shot network quantization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12339–12348 (2022)