

Accurate Zero-shot Quantization via Hierarchical Teacher-Assistant Distillation

Wonjin Cho[✉], Jeongin Yun[✉], and U Kang[✉]

Seoul National University, Seoul, South Korea
{chowonjin0627, yji00828, ukang}@snu.ac.kr

Abstract. How can we bridge the numerical chasm between full-precision CNN models and highly quantized targets when original training data are unavailable? As the demand for deploying deep neural networks on resource-constrained edge hardware grows, achieving extreme model compression without compromising the privacy of training data has become a critical challenge. However, direct transitions to ultra-low-bit regimes introduce extreme optimization instability and activation collapse, where high quantization noise causes deep layers to lose their discriminative capacity and representational power.

In this paper, we propose ZEST (ZSQ Enhancement via Stepping-stone Transition), an accurate zero-shot quantization framework for CNNs that bridges the precision gap between full-precision models and ultra-low-bit targets through a unified two-stage distillation and reconstruction pipeline. Rather than attempting a direct and unstable jump from full precision to an ultra-low bit quantized model, ZEST introduces a stable intermediate bit stage to tame the numerical distribution and regularize the model. Furthermore, we propose a two-phase distillation strategy that initializes high-fidelity semantic skeletons from the full-precision model and subsequently recalibrates them through an intermediate bit stage to ensure calibration data are numerically aligned with quantized-domain constraints. Experimental results demonstrate that this progressive transition leads to better distillation, achieving 70.50% top-1 accuracy on ResNet-18 at $W4A4$ precision and showing higher performance compared to existing SOTA methods in various models.

Keywords: CNN · Zero-shot Quantization · Knowledge Distillation

1 Introduction

Given a pretrained full-precision model without its training data, how can we compress it while preserving the original model’s accuracy and representational logic? This problem is critical for deploying high-performance AI on resource-constrained edge devices where privacy or proprietary restrictions prevent the sharing of raw data. Solving this problem enables secure, high-speed model acceleration for third-party service providers and mobile hardware manufacturers. Quantization is especially effective at maintaining model performance in data-free scenario [7, 10, 17, 19, 26, 27, 31] compared to other model compression

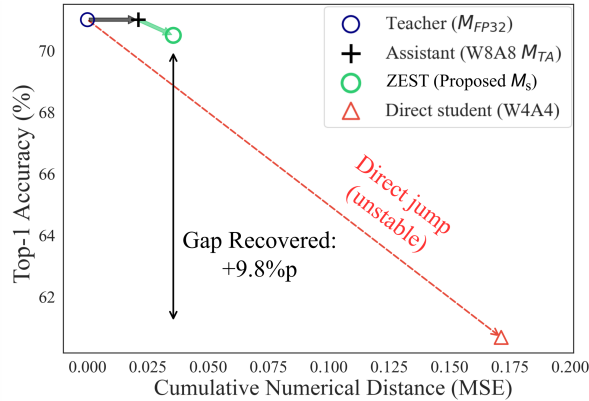


Fig. 1: The hierarchical advantage. ZEST bridges the representation gap by proposing an intermediate assistant model, achieving superior accuracy. In contrast, the direct method exploits synthetic images and knowledge distillation exclusively from the full-precision teacher; this creates a significant optimization gap that results in substantial accuracy loss.

techniques [8, 20], such as low-rank decomposition [11, 30, 34], knowledge distillation [4, 13, 15, 16, 36, 39], architectural design [12, 43], and pruning [28, 29, 37, 40].

Despite their effectiveness, extending zero-shot quantization to ultra-low-bit regimes remains challenging: a direct full-precision to ultra-low-bit quantization induces a severe representation gap and optimization instability. Most existing zero-shot quantization methods [14, 18, 22, 23, 42, 44] perform calibration and distillation in a single step using synthetic data derived from BatchNorm statistics. However, under extreme quantization noise, this one-step calibration is often insufficient, frequently triggering “activation collapse” in which signal variance diminishes in deeper layers. As a result, when the capacity gap between the full-precision model and the quantized model becomes too large, the accuracy of the quantized model degrades catastrophically [5, 32]. Fig. 1 shows that direct quantization creates a significant representation gap, which leads to substantial accuracy degradation compared to ZEST.

In this paper, we propose ZEST, an accurate zero-shot quantization framework for CNNs that enables a stable transition from full-precision models to ultra-low-bit quantized models through a unified two-stage distillation and reconstruction pipeline. Instead of a single-step leap, ZEST establishes a stable foundation by transitioning the full-precision model to an intermediate bit-width. This intermediate anchor serves as a proximal teacher, taming the distribution of feature maps and providing a smoother optimization landscape for the final ultra-low-bit reconstruction.

The primary contributions of this work are as follows:

- **Optimization stability analysis.** We analyze the impact of quantization noise on gradient variance and demonstrate that an intermediate-precision proxy acts as a structural noise filter. This hierarchical transition reduces the

gradient bias and optimization instability that often plagues direct ultra-low-bit reconstruction.

- **Algorithm.** We propose a unified framework for generating numerically stable synthetic data that bridges the precision gap in zero-shot quantization. This is necessary to ensure that the synthetic data inherently accounts for quantized-domain constraints, such as clipping effects and mean shifts, which are often ignored by traditional direct methods. ZEST achieves this through refining synthetic images to reflect quantized-domain constraints—such as clipping effects—under the guidance of a Teacher Assistant (TA).
- **Empirical validation.** We demonstrate that an intermediate-precision TA acts as a stable numerical anchor for ultra-low-bit targets. Experimental results show that the ZEST recovers the majority of accuracy lost due to quantization, bringing the *W4A4* ResNet-18 to within 0.51 percentage points of the full-precision baseline.

The rest of the paper is organized as follows: Section 2 describes preliminaries, Section 3 details the proposed ZEST framework, Section 4 presents experimental results, Section 5 reviews related work, and Section 6 concludes the paper. To enhance the readability of this paper, we list the symbols frequently used in this paper in the Supplementary Materials [1]. We provide the source code used in this paper in <https://github.com/snudm-starlab/ZEST/tree/main>.

2 Preliminaries

2.1 Problem Definition

Zero-shot quantization is an effective approach to address the model compression problem. We formally define the problem of zero-shot quantization for ultra-low bit-widths as follows:

Problem 1 (Zero-shot quantization). **Given** a pretrained full-precision teacher model $M_{teacher}$ and the absence of the original training data $\mathcal{D}$, **quantize** a student model M_S whose bit-width k is constrained to ultra-low regimes.

2.2 Architecture of the Convolutional Neural Network (CNN)

A CNN consists of a sequence of convolutional layers that extract features using learnable filters. As the depth of the network increases, feature representations progressively shift from local low-level patterns to higher-level semantic representations. Standard CNN architectures (e.g., ResNet [9]) often group layers into blocks, where each block functions as a modular unit containing multiple layers; residual networks additionally include shortcut connections within these blocks. Our proposed ZEST exploits these blocks as the primary structural units for quantization and distillation.

2.3 Uniform Quantization

We focus on uniform symmetric quantization, which is hardware-friendly and widely supported. For a given tensor x (weights or activations), the quantization process maps floating-point values to k -bit integers. The quantization function $Q(x)$ is defined as:

$$Q(x) = \text{clamp} \left(\left\lfloor \frac{x}{s} \right\rfloor, -2^{k-1}, 2^{k-1} - 1 \right), \quad (1)$$

where s is the scaling factor and $\lfloor \cdot \rfloor$ denotes the nearest rounding operation. The scaling factor s is determined by the dynamic range of the tensor, calculated as $s = \frac{\max(|x|)}{2^{k-1}-1}$. The de-quantization process approximates the original value:

$$\hat{x} = s \cdot Q(x).$$

2.4 Knowledge Distillation (KD)

Standard KD transfers the knowledge from a teacher model $M_{teacher}$ to a student model M_S by minimizing the Kullback-Leibler (KL) divergence between their output probability distributions. The loss function $\mathcal{L}_{KD}$ is defined as:

$$\mathcal{L}_{KD} = \sum_i p_i^T(\tau) \log \left(\frac{p_i^T(\tau)}{p_i^S(\tau)} \right),$$

where $p_i = \frac{\exp(z_i/\tau)}{\sum_j \exp(z_j/\tau)}$ are the softened probabilities with temperature τ ; z_i and z_j represent the i -th and j -th elements of the logit vector. $p_i^T(\tau)$ and $p_i^S(\tau)$ denote the softened output probabilities of the teacher model $M_{teacher}$ and the student model M_S , respectively. In existing zero-shot quantization, KD is typically performed via end-to-end reconstruction as shown above, where the entire student network is optimized simultaneously to match the teacher’s final output.

3 Proposed Method

We propose ZEST (ZSQ Enhancement via Stepping-stone Transition), an accurate zero-shot quantization framework for CNN models. These are three main challenges that must be tackled:

- C1. Optimization instability from quantization noise.** A direct jump from full-precision to ultra-low-bit regimes used in previous methods injects excessive quantization noise that biases gradients and prevents stable optimization toward meaningful local minima. How can we reduce the excessive quantization noise that severely biases gradients caused by direct quantization without the presence of real data?

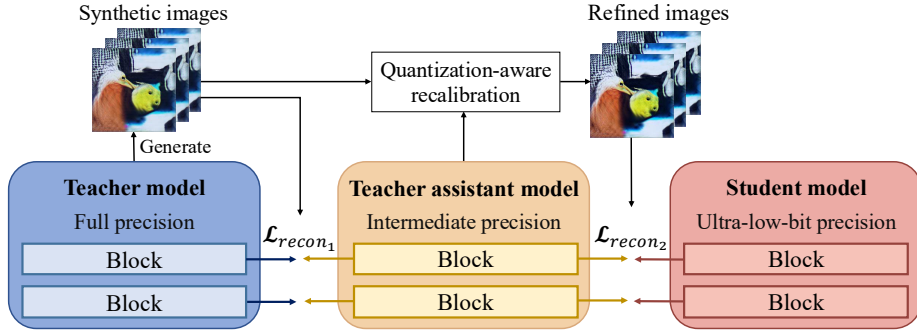


Fig. 2: Hierarchical data-free quantization pipeline. The framework exploits a TA with intermediate bit-width to bridge the numerical gap between full-precision and low-bit models via two-phase distillation.

- C2. Numerical domain mismatch of synthetic calibration data.** Despite capturing high-level semantics, conventional synthesis methods fail to reflect the specific numerical clipping thresholds of quantized models, resulting in domain mismatch that exacerbates error accumulation. How can we synthesize calibration data that preserve the rich semantic skeleton while reflecting the unique numerical constraints of the quantized domain?
- C3. Activation distribution collapse in deep layers.** Standard distillation frameworks fail to calibrate the vast representational gap between a high-precision teacher and an ultra-low-bit student, leading to “activation collapse” in deeper layers where discriminative signal variance vanishes and representational capacity is lost. How can we prevent “activation collapse” in the deeper layers of a network, where discriminative signal variance vanishes due to aggressive bit-width reduction?

To address these challenges, we propose the following three core ideas:

- I1. Numerical stepping stone.** We propose a single TA at an intermediate precision to provide a proximal optimization target and a smoother loss surface for the student.
- I2. Quantization-aware image recalibration.** We refine synthetic images by referencing the TA to stabilize optimization, pre-aligning the calibration data with quantized-domain constraints.
- I3. Block-wise reconstruction.** We propose a block-wise reconstruction pipeline guided by TA to preserve signal variance and align feature magnitudes with low-bit student.

Fig. 2 illustrates the overall pipeline of ZEST. Our framework uses three models at different bit-widths: a full-precision teacher, an intermediate-precision TA, and an ultra-low-bit student. Algorithm 1 details the procedure for bridging the numerical gap from the full-precision model to the ultra-low-bit student. The procedure follows a two-stage pipeline, consisting of TA fine-tuning (lines 3–14) followed by student fine-tuning (lines 15–23). In the first stage, we aim to train an accurate TA that serves as a numerical bridge for stable data-free reconstruction

Algorithm 1 Overall framework of ZEST

```

1: Input: Full-precision model  $M_{teacher}$ , target bit-width  $k$ , number  $T$  of batches,
   step-size  $\eta$ , distill iteration  $N$ , and recalibration iteration  $U$ 
2: Output: Optimized ultra-low-bit student model  $M_S$ 
   /* Fine-tuning TA model (Section 3.2) */
3:  $X_{teacher} \leftarrow \emptyset$ 
4: Initialize a generator  $G$ 
5: repeat
6:   Update  $G$  with an alignment loss  $\mathcal{L}_{BN}$  in Eq. (3)
7: until  $N$  times
8: Freeze  $G$ 
9: repeat
10:   Generate a batch  $\mathbf{x}_{teacher}$  of synthetic images from the generator  $G$ 
11:    $X_{teacher} \leftarrow X_{teacher} \cup \{\mathbf{x}_{teacher}\}$ 
12: until  $T$  times
13: Quantize  $M_{teacher}$  to intermediate precision to initialize  $M_{TA}$ 
14: Fine-tune  $M_{TA}$  via block-wise optimization based on  $X_{teacher}$ 
   /* Fine-tuning student model (Section 3.3) */
15:  $X_{TA} \leftarrow X_{teacher}$ 
16: for each batch  $\mathbf{x}_{TA} \in X_{TA}$  do
17:   repeat
18:     Calculate the refinement loss  $\mathcal{L}_{TA}$  in Eq. (5)
19:     Refine images:  $\mathbf{x}_{TA} \leftarrow \mathbf{x}_{TA} - \eta \cdot \nabla_{\mathbf{x}_{TA}} \mathcal{L}_{TA}$ 
20:   until  $U$  times
21: end for
22: Quantize  $M_{teacher}$  to target low-bit precision  $k$  to initialize student model  $M_S$ 
23: Fine-tune  $M_S$  block-wise based on  $X_{TA}$  and  $M_{TA}$ 
24: return  $M_S$ 

```

of the ultra-low-bit student. We synthesize calibration images matched to the teacher’s statistics (lines 3–12), then quantize and fine-tune the TA to retain the teacher’s semantics at intermediate precision (lines 13–14). In the second stage, we aim to train the student effectively by leveraging the TA as a quantization-aware supervisor. Specifically, we refine the synthetic images via the TA to better reflect quantized-domain effects (lines 15–21), and then fine-tune the student at the target bit-width with the refined images under TA guidance (lines 22–23).

3.1 TA Bridge in Distillation-Numerical Stepping Stone

Our goal is to achieve stable and accurate optimization for ultra-low-bit student models in a data-free regime. The main challenge to achieve this goal is the excessive quantization noise injected during the direct transition from a full-precision teacher to an ultra-low-bit student. This massive numerical gap generates biased gradients in the absence of real data, triggering optimization instability and preventing the student from reaching a meaningful local minimum. Our idea to solve this challenge is the TA bridge, which proposes an intermediate bit-width TA model to provide a numerical stepping stone for the student. Diverging from previous methods that use intermediate layer counts [24,35], we tailor our approach

for the quantization regime while maintaining the same number of layers as the teacher and student models. The TA model learns from the teacher model and teaches the student model, which we further discuss in Section 3.3. We determine the precision of the TA bridge model according to the following hierarchy:

$$b(M_S) < b(M_{TA}) < b(M_{teacher}) \quad (2)$$

where $b(M)$ denotes the bit-width of model M ; M_S , M_{TA} , and $M_{teacher}$ denote the student, teacher-assistant, and teacher models, respectively. We find that a single TA bridge yields the best student performance in a zero-shot setting. We validate this conclusion empirically by varying the number of TAs: as the number of TAs increases, student performance consistently degrades, with a single TA still outperforming the variant without any TA. We provide detailed experimental results in the Supplementary Materials [1].

The primary advantage of the TA bridge is that it acts as a structural noise filter by providing a proximal optimization target. Specifically, the TA keeps the student weights within a stable region of the loss surface during optimization by serving as a high-fidelity anchor that is numerically closer to the final ultra-low-bit student model. We also verify this effect empirically: the intermediate TA reduces the variance of both gradient and reconstruction loss during the student optimization phase (see Section 4.4). These results suggest that TA-based supervision suppresses high-variance numerical noise inherent in direct distillation, allowing the student gradients to focus on essential semantic features rather than absorbing noise-induced bias.

3.2 Fine-tuning a TA Model

Our goal is to create a numerical anchor for an ultra-low-bit student model in a data-free regime. The main challenge to achieve this goal is the absence of training data when provided only with the pretrained model $M_{teacher}$. Our main idea to solve this challenge is to create synthetic data by capturing the global semantic manifold of the unknown original data from the full precision model $M_{teacher}$. We then initialize the TA model M_{TA} by quantizing the base model $M_{teacher}$ according to Eq. (1) at a bit-width satisfying Eq. (2) and fine-tune the TA with the synthetic data. Following the standard ZSQ paradigm, we generate synthetic images $\mathbf{x}_{teacher}$ with a generator model G which maps random noise $\mathbf{z}$ conditioned on uniformly sampled label y to images:

$$\mathbf{x}_{teacher} = G(\mathbf{z}, y), \quad \mathbf{z} \sim \mathcal{N}(0, 1).$$

We train the generator G by aligning the batch-normalization statistics induced by synthetic images in the full precision model with the pre-stored statistics. We formally define the alignment loss $\mathcal{L}_{BN}$ as follows:

$$\mathcal{L}_{BN} = \sum_l (\|\mu_l^s - \mu_l\|_2^2 + \|\sigma_l^s - \sigma_l\|_2^2), \quad (3)$$

where μ_l^s and σ_l^s denote the mean and variance induced by the synthetic images at the l -th batch-normalization layer of the full-precision model $M_{teacher}$; μ_l and σ_l denote the corresponding pre-stored mean and variance of $M_{teacher}$ in that layer.

With the generated synthetic images $\mathbf{x}_{teacher}$, we fine-tune the quantized TA via a block-wise reconstruction pipeline sequentially from the first to the last block to prevent error accumulation before it reaches deeper layers. We independently fine-tune each block by matching the block output of a *supervisor* model to that of a *learner* model. Specifically, for the b -th block, we optimize the TA model M_{TA} with the reconstruction loss $\mathcal{L}_{recon,1}^b$ on the base synthetic images:

$$\mathcal{L}_{recon_1}^b = \mathbb{E}_{\mathbf{x}_{teacher}} \|A_{M_{teacher}}^b(\mathbf{x}_{teacher}) - A_{M_{TA}}^b(\mathbf{x}_{teacher})\|_2^2, \quad (4)$$

where $A_M^b(\mathbf{x})$ denotes the output activation of the b -th block of model M given input $\mathbf{x}$.

3.3 Fine-tuning a Student Model

Our goal is to achieve stable and accurate optimization for ultra-low-bit student models by learning from the TA model. Existing data-free distillation methods typically transfer knowledge from a teacher to a student using base synthetic images $\mathbf{x}_{teacher}$ [14, 18, 22, 44]. However, these images are optimized for full-precision statistics and fail to account for the unique numerical clipping and distribution shifts inherent in quantized networks. Our main idea to solve this challenge is to refine the base synthetic images $\mathbf{x}_{teacher}$ into TA-specific images $\mathbf{x}_{TA}$ that reflect the quantized internal distribution of the TA model, thereby providing a more compatible optimization target for the student.

We initialize $\mathbf{x}_{TA}$ with $\mathbf{x}_{teacher}$ and directly update the pixels of synthetic images $\mathbf{x}_{TA}$ with a refinement loss $\mathcal{L}_{TA}$. We define the update rule and the loss as follows:

$$\begin{aligned} \mathbf{x}_{TA} &\leftarrow \mathbf{x}_{TA} - \eta \cdot \nabla_{\mathbf{x}_{TA}} \mathcal{L}_{TA}, \\ \mathcal{L}_{TA} &= \sum_l \left(\|\mu_l^t - \mu_l^{M_{TA}}\|_2^2 + \|\sigma_l^t - \sigma_l^{M_{TA}}\|_2^2 \right), \end{aligned} \quad (5)$$

where μ_l^t and σ_l^t denote the mean and variance induced by the synthetic images at the l -th batch-normalization layer of the TA model M_{TA} ; $\mu_l^{M_{TA}}$ and $\sigma_l^{M_{TA}}$ denote the corresponding pre-stored mean and variance in that layer; η denotes the step size for gradient updates. We perform gradient-descent updates U times to obtain the final refined images. This update matches the re-calibrated BN statistics (μ^{TA}, σ^{TA}) that shift by quantization bias [38, 41], with small number of iterations by forcing $U < N$, where N denotes the number of iteration for updating the generator G for $\mathbf{x}_{teacher}$.

Directly updating base synthetic images is superior to generating images from a quantized model from scratch. The images require only a subtle numerical “shift” to account for clipping boundaries since the full-precision and intermediate models share global topology. The proposed U -step update therefore

achieves “quantized-domain alignment” while preserving the variance necessary for the reconstruction of a stable ultra-low-bit model.

Following the image refinement, we initialize the student model M_S by quantizing the base model $M_{teacher}$ according to Eq. (1) to the target ultra-low bit-width as shown in Eq. (2). We then fine-tune the quantized student model via a block-wise reconstruction pipeline sequentially from the first to the last block to prevent error accumulation before it reaches deeper layers. Specifically, for the b -th block, we optimize the student model M_S with the reconstruction loss $\mathcal{L}_{recon,2}^b$ on the refined images:

$$\mathcal{L}_{recon,2}^b = \mathbb{E}_{\mathbf{x}_{TA}} \|A_{M_{TA}}^b(\mathbf{x}_{TA}) - A_{M_S}^b(\mathbf{x}_{TA})\|_2^2, \quad (6)$$

where $A_M^b(\mathbf{x})$ is the activation of the b -th block of model M given input $\mathbf{x}$.

4 Experiments

To evaluate the efficacy of ZEST, we conduct extensive experiments on large-scale computer vision benchmarks. The experiments address the following research questions in detail:

- Q1. Accuracy (Section 4.2).** How does the TA-guided framework perform against the state-of-the-art baselines in ultra-low-bit regimes?
- Q2. TA selection (Section 4.3).** What is the optimal TA precision for training ultra-low-bit quantized student models?
- Q3. Optimization analysis (Section 4.4).** What is the impact of the TA bridge on optimization stability?
- Q4. Ablation analysis (Section 4.5).** What is the individual impact of each proposed component on final student fidelity?

4.1 Experimental Settings

Benchmarks and architectures. We evaluate the performance of ZEST framework on the ImageNet (ILSVRC12) dataset [21], which consists of 1.2M training images and 50k validation images. The proposed data-free pipeline performs quantization without access to real training data and learns only from the pretrained full-precision model. We validate ZEST across diverse architectures, including ResNet-18, ResNet-50, and the more challenging MobileNetV2 [33]. We compare ZEST against leading data-free quantization baselines: ZeroQ [2], GDFQ [38], AdaDFQ [32], Qimera [5], Genie [14], TexQ [3], and SynQ [18].

Quantization and generator details. The proposed framework is developed using PyTorch 23.07 with CUDA 12.1, executed on an NVIDIA A100 (40GB) GPU. For the intermediate Teacher Assistant (TA), we exploit standard min-max quantization at $W8A8$ precision. For the target ultra-low-bit students ($W4A4$ and $W3A3$), we exploit Learnable Step Size Quantization (LSQ) [6]. We follow

Table 1: Top-1 accuracy comparison on ImageNet-1K (W4A4). ZEST demonstrates competitive or superior performance against generative baselines.

Method	Bits (W/A)	ResNet-18	ResNet-50	MobileNetV2
Full Prec.	32/32	71.01%	76.63%	71.88%
ZeroQ [2]	4/4	60.68%	-	59.39%
GDFQ [38]	4/4	60.60%	-	-
IntraQ [44]	4/4	66.47%	-	69.10%
AdaDFQ [32]	4/4	66.53%	68.38%	65.41%
Qimera [5]	4/4	63.84%	66.25%	61.62%
Genie [14]	4/4	69.66%	75.59%	68.38%
TexQ [3]	4/4	67.73%	70.72%	67.07%
GenQ [23]	4/4	69.77%	75.50%	68.96%
SynQ [18]	4/4	67.90%	71.05%	67.27%
ZEST (proposed)	4/4	70.50%	76.32%	70.71%

the standard PTQ protocols by maintaining the first layer at 8-bit to preserve sensitivity at the network boundaries. The generator follows a standard ResNet-style architecture, optimized for matching BN statistics.

Implementation specifics. The proposed hierarchical pipeline fine-tunes the TA model, then the student model. When fine-tuning the TA model quantized to W8A8, we generate 1,024 synthetic samples over 4,000 iterations with the teacher model. When fine-tuning a student model, the TA model acts as the proximal teacher. We update synthetic images generated from the teacher model for 500 iterations with the TA model to capture quantized-domain features. The block-wise reconstruction is performed over 20,000 iterations per block with a batch size of 32. In all tables, the best performance is highlighted in bold.

4.2 Performance on Low-bit Regimes

We evaluate the performance of the proposed hierarchical framework across various low-bit configurations to determine if ZEST effectively reclaims the accuracy lost due to the aggressive reduction of bit-width. The evaluation primarily focuses on the 4-bit (W4A4) and 3-bit (W3A3) regimes, which represent the current frontiers of zero-shot quantization in ultra-low-bit. Comparing the proposed ZEST against generative and diffusion-based baselines demonstrates that maintaining numerical stability through an intermediate anchor is as critical as the quality of the synthetic calibration data itself.

Analysis on 4-bit quantization. Table 1 summarizes the Top-1 accuracy of ZEST compared to the recent state-of-the-art (SOTA) data-free quantization methods. ZEST quantizes all layers except the first, which maintains 8-bit precision for weights following Genie [14] and GenQ [23]. Table 1 shows that ZEST significantly outperforms earlier generative methods. On ResNet-18, we achieve

Table 2: Top-1 accuracy in the ultra-low-bit regime ($W3A3$) on ImageNet-1K.

Method	Bits	ResNet-18	ResNet-50	MobileNetV2
GDFQ [38]	3/3	20.23%	0.31%	1.46%
AdaDFQ [32]	3/3	38.10%	17.63%	28.99%
Genie [14]	3/3	66.16%	71.61%	57.54%
TexQ [3]	3/3	50.28%	25.27%	32.80%
GenQ [23]	3/3	68.18%	73.99%	59.15%
SynQ [18]	3/3	52.02 %	26.89 %	34.21 %
ZEST (proposed)	3/3	68.45%	75.28%	67.74%

Table 3: Ablation of TA bit-width. Ablation on ResNet-18 model quantized to $W4A4$ with ImageNet-1K validation set

TA Precision	Student Top-1 (%)
$W16A16$ (FP)	69.15
$W8A8$ (Integer)	70.50
$W6A6$ (Integer)	70.34
$W4A4$ (Student)	69.19

70.50%, surpassing previous SOTA methods. This validates that the intermediate TA stage mitigates the reconstruction error inherent in aggressive quantization.

Performance on 3-bit quantization model. To further test the robustness of the TA bridge, we push the quantization limits to $W3A3$. As shown in Table 2, the accuracy degradation in direct quantization is significantly higher than that of ZEST. Traditional methods such as GDFQ and AdaDFQ collapse significantly. Other methods, including Genie and GenQ, preserve their resilience, but ZEST maintains a much higher accuracy of 68.45%, demonstrating that the hierarchical distillation path is essential for preserving feature representation when bit-width is extremely limited.

4.3 Analysis of TA Selection

We investigate the critical role of the TA’s precision as a proximal optimization target and summarize the results in Table 3. Choosing the correct anchor bit-width involves a trade-off between preserving the semantic integrity of the teacher and minimizing the proximal gap to the low-bit student. We investigate the impact of the TA’s precision on final student performance, as summarized in Table 3. The results reveal a non-monotonic relationship between anchor precision and student accuracy. This validates the trade-off between semantic fidelity and numerical proximity.

The high precision anchor ($W16A16$) yields sub-optimal student performance (69.15%) even though it maintains near-perfect semantic fidelity. This is because

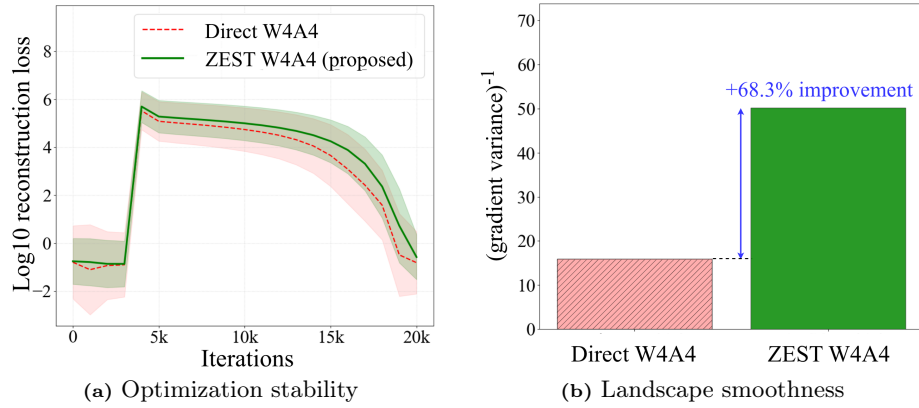


Fig. 3: Comparison of reconstruction loss convergence. (a) ZEST (green) demonstrates superior optimization stability with tighter variance across iterations compared to the direct baseline (red). (b) ZEST achieves a +68.3 percentage point improvement in the smoothness index ($1/\text{gradient variance}$), indicating a significantly flatter and more stable loss landscape for quantization.

a floating-point teacher fails to model the clipping and rounding noise that the student will inevitably encounter, leaving the “precision gap” largely unbridged. Conversely, decreasing the TA precision to W6A6 leads to an accuracy drop (70.34%) due to semantic drift, where the quantization error in the teacher itself begins to degrade the quality of the guidance. The W8A8 configuration achieves the highest fidelity (70.50%) by acting as the ideal numerical bridge. At 8-bit, the model retains sufficient representational capacity to guide the student’s logic while introducing the necessary quantized-domain priors—such as limited dynamic range—required for stable ultra-low-bit reconstruction.

4.4 Empirical Analysis of Optimization Stability

We conduct an in-depth empirical analysis of the training dynamics of the ultra-low-bit student (W4A4) to validate the claim in Sec. 3.1 that the TA bridge acts as a proximal optimization target to reduce optimization volatility. We compare ZEST against direct distillation from the full-precision teacher. The results are visualized in Fig. 3, which shows how the TA bridge stabilizes the optimization process for ultra-low-bit models. Fig. 3(a) reveals that ZEST maintains a significantly tighter variance corridor compared to the direct distillation that suffers from substantial fluctuations in reconstruction loss—indicated by the wide shaded variance region. This stabilization is further quantified in Fig. 3(b) through the smoothness index, which measures the inverse of the gradient variance; ZEST achieves +68.3 percentage point improvement in landscape smoothness compared to the direct baseline. ZEST effectively mitigates the “divergence trap” caused by excessive quantization noise by utilizing the TA as a proximal anchor, resulting in less noisy gradients and a more consistent traversal toward a meaningful local minimum. We additionally evaluate the variance across five

Table 4: Ablation study of proposed ZEST components. Results are reported for ResNet-18 on ImageNet-1K at W4A4 precision.

Method	Top-1 Accuracy(%)
ZEST-T (w/o TA)	70.23
ZEST-R (w/o refinement)	69.94
ZEST-B (w/o block-wise)	69.99
ZEST (proposed)	70.50

independent training runs; the W4A4 student on ResNet-18 achieves a mean Top-1 accuracy of $70.42\% \pm 0.08\%$.

4.5 Ablation Study

To evaluate the proposed ideas, we conduct a comprehensive ablation study to isolate the performance gains attributable to each module of ZEST framework. The results highlight the non-linear relationship between the proposed components. Table 4 compares the full model (ZEST) with its variants that exclude each of the three proposed ideas: (1) ZEST-T which removes the TA bridge in Section 3.1, (2) ZEST-R which replaces the refined synthetic images $\mathbf{x}_{TA}$ in Section 3.3 with the base synthetic images $\mathbf{x}_{teacher}$ from the teacher model $M_{teacher}$, and (3) ZEST-B which replaces the block-wise reconstruction in Eq. (4) and Eq. (6) with end-to-end reconstruction over the final logits of the supervisor and the learner. The results show that removing any component degrades accuracy, demonstrating that all components are essential for the model’s performance. Notably, the exclusion of image refinement (ZEST-R) leads to a significant accuracy drop despite having the TA architecture and block-wise reconstruction. This reinforces the proposed “quantized-domain alignment” hypothesis: even a high-capacity teacher is ineffective if the calibration data are optimized for a mismatched numerical domain.

5 Related Works

5.1 Zero-Shot Quantization (ZSQ)

Zero-Shot Quantization (ZSQ) is a critical research frontier for deploying deep neural networks in environments where original training data are inaccessible due to privacy mandates or proprietary restrictions. Existing ZSQ methodologies are classified into synthesis-free and synthesis-based approaches.

Synthesis-free methods. Early works, such as the seminal DFQ [25], leverage analytical techniques to equalize weight ranges across channels and utilize bias correction to mitigate mean shifts in activation distributions. While efficient for 8-bit quantization, they fail to model the non-linear distortions inherent in ultra-low bit-widths. ZEST employs a dynamic synthesis framework that adaptively manages the increased noise floor of low-precision regimes.

BNS-guided data synthesis. To address the limitations of analytic methods, synthesis-based approaches reconstruct calibration data by inverting the knowledge encapsulated within a pretrained teacher model. ZeroQ [2] pioneers the optimization of input noise to match the Batch Normalization (BN) statistics of the teacher. The objective is to minimize the discrepancy between the synthetic distribution $\mathcal{D}_{syn}$ and the BN statistics across all L layers. This ensures that synthetic samples elicit a similar neural response to the original data manifold. GDFQ [38] extends this paradigm by training a generator network $\mathcal{G}$ to synthesize diverse samples, incorporating cross-entropy loss to enforce semantic consistency. Other research papers improve this idea [3, 14, 32]. However, these methods frequently encounter *mode collapse*, where the generator yields redundant samples that satisfy global statistics but fail to represent the local intricacies of the feature space. While these works rely on a direct distillation from a full-precision teacher, we propose a stepping stone precision transition to prevent such optimization failures and ensure high-fidelity data synthesis.

5.2 Knowledge Distillation via Teacher Assistant

The concept of using an intermediate model to improve distillation is formalized by Mirzadeh et al [24]. They observe that when the gap in capacity (e.g., depth, width) between teacher and student is too large, the student fails to track the teacher’s complex decision boundary. By introducing a TA of intermediate capacity, the distillation process $T \rightarrow TA \rightarrow S$ yields a lower total error than direct $T \rightarrow S$ distillation. While knowledge distillation via teacher assistant is originally proposed for architectural gaps (e.g., ResNet101 to ResNet18), our work is the first to generalize this concept to the numerical precision domain, addressing the unique noise and clipping challenges of ultra-low-bit quantization.

6 Conclusion

We propose ZEST, a hierarchical framework for zero-shot quantization that bridges the precision gap in ultra-low-bit networks. By utilizing a *model with intermediate bit-width* as a numerical anchor, ZEST effectively regularizes the feature distributions for 4-bit and 3-bit targets. ZEST demonstrates that semantic initialization from a full-precision model followed by efficient TA refinement prevents the distributional collapse prevalent in data-free methods. Experimental results on ImageNet-1K validate ZEST, with ResNet-18 achieving 70.50% Top-1 accuracy at W4A4. Our analysis confirms that a single intermediate anchor is optimal for minimizing quantization error while maintaining semantic fidelity. Future work will explore the generalizability of this TA-guided pipeline to Large Language Models (LLMs) and Vision Transformers (ViTs), where outlier activations present significant challenges for standard post-training quantization.

Acknowledgement

This work was supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) [No.RS-2020-II200894, Flexible and Efficient Model Compression Method for Various Applications and Environments], [No.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)], [No.RS-2025-25442338, AI Star Fellowship Support Program(Seoul National Univ.)], [No.RS-2024-00509257, Global AI Frontier Lab], and [No.2022-0-00641, XVoice: Multi-Modal Voice Meta Learning]. This work was also supported by Youlchon Foundation.

References

1. Authors, A.: Supplementary materials (submitted). In: ECCV Submission (2026)
2. Cai, Y., et al.: Zeroq: A novel zero shot quantization framework. In: CVPR (2020)
3. Chen, X., Wang, Y., Yan, R., Liu, Y., Guan, T., He, Y.: Texq: Zero-shot network quantization with texture feature distribution calibration. In: NeurIPS (2023)
4. Cho, I., Kang, U.: Pea-kd: Parameter-efficient and accurate knowledge distillation on bert. PLOS ONE **17**(2), 1–12 (02 2022)
5. Choi, K., et al.: Qimera: Exploiting generating model for zero-shot quantization. In: NeurIPS (2021)
6. Esser, S.K., McKinstry, J.L., Bablani, D., Appuswamy, R., Modha, D.S.: Learned step size quantization. In: ICLR (2020)
7. Gholami, A., Kim, S., Dong, Z., Yao, Z., Mahoney, M.W., Keutzer, K.: A survey of quantization methods for efficient neural network inference. In: Low-power computer vision, pp. 291–326. Chapman and Hall/CRC (2022)
8. Han, S., Mao, H., Dally, W.J.: Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In: ICLR (2016)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
10. Hubara, I., et al.: Quantized neural networks: Training neural networks with low precision weights and activations. In: JMLR (2018)
11. Jaderberg, M., Vedaldi, A., Zisserman, A.: Speeding up convolutional neural networks with low rank expansions (2014)
12. Jang, J., Quan, C., Lee, H.D., Kang, U.: Falcon: lightweight and accurate convolution based on depthwise separable convolution. Knowl. Inf. Syst. **65**(5), 2225–2249 (2023)
13. Jeon, H., Park, S., Kim, J.G., Kang, U.: Pet: Parameter-efficient knowledge distillation on transformer. PLOS ONE **18**, 1–21 (07 2023)
14. Jeon, Y., Lee, C., Kim, H.y.: Genie: Show me the data for quantization. In: CVPR (2023)
15. Jiang, Y., Feng, C., Zhang, F., Bull, D.: Mtkd: Multi-teacher knowledge distillation for image super-resolution. In: ECCV (2024)
16. Kim, J., Jung, J., Kang, U.: Compressing deep graph convolution network with multi-staged knowledge distillation. PLOS ONE **16**, 1–18 (08 2021)
17. Kim, M., Choi, J., Yang, H., Kim, J., Song, J., Kang, U.: Prune-then-quantize or quantize-then-prune? understanding the impact of compression order in joint model compression. In: ICLR (2026)

18. Kim, M., Kim, J., Kang, U.: Synq: Accurate zero-shot quantization by synthesis-aware fine-tuning. In: ICLR (2025)
19. Kim, M., Lee, J., Kim, J., Yun, J., Kwon, Y., Kang, U.: Lampq: Towards accurate layer-wise mixed precision quantization for vision transformers. In: AAAI (2026)
20. Kim, M., et al.: Zero-shot quantization: A comprehensive survey. In: IJCAI (2025)
21. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NeurIPS (2012)
22. 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: CVPR (2023)
23. Li, Y., Kim, Y., Lee, D., Kundu, S., Panda, P.: Genq: Quantization in low data regimes with generative synthetic data. In: ECCV (2024)
24. Mirzadeh, S.I., Farajtabar, M., Li, A., Levine, N., Matsukawa, A., Ghasemzadeh, H.: Improved knowledge distillation via teacher assistant. In: AAAI (2020)
25. Nagel, M., et al.: Data-free quantization through weight equalization and bias correction. In: ICCV (2019)
26. Nagel, M., et al.: Up or down? adaptive rounding for post-training quantization. In: ICML (2020)
27. Park, S., Kim, M., Kang, U.: Unifying uniform and binary-coding quantization for accurate compression of large language models. In: ACL (2025)
28. Park, S., Lee, S., Kim, J., Lee, J., Jo, H., Kang, U.: Accurate sublayer pruning for large language models by exploiting latency and tunability information. In: IJCAI (2025)
29. Park, S., et al.: Accurate retraining-free pruning for pretrained encoder-based language models. In: ICLR (2024)
30. Phan, A.H., Sobolev, K., Sozykin, K., Ermilov, D., Gusak, J., Tichavsky, P., Glukhov, V., Oseledets, I., Cichocki, A.: Stable low-rank tensor decomposition for compression of convolutional neural network. In: ECCV (2020)
31. Piao, T., Cho, I., Kang, U.: Sensimix: Sensitivity-aware 8-bit index & 1-bit value mixed precision quantization for bert compression. PloS one **17**(4), e0265621 (2022)
32. Qian, B., et al.: Adaptive data-free quantization. In: CVPR (2023)
33. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: CVPR (2018)
34. Schotthöfer, S., Zangrando, E., Kusch, J., Ceruti, G., Tudisco, F.: Low-rank lottery tickets: finding efficient low-rank neural networks via matrix differential equations. In: NeurIPS (2022)
35. Son, W., Na, J., Choi, J., Hwang, W.: Densely guided knowledge distillation using multiple teacher assistants. In: ICCV (2021)
36. Wang, Y., Cheng, L., Duan, M., Wang, Y., Feng, Z., Kong, S.: Improving knowledge distillation via regularizing feature norm and direction. In: ECCV (2024)
37. Wang, Z., Li, C., Wang, X.: Convolutional neural network pruning with structural redundancy reduction (2021)
38. Xu, S., et al.: Generative low-bitwidth data free quantization. In: ECCV (2020)
39. Yang, C., Zhou, H., An, Z., Jiang, X., Xu, Y., Zhang, Q.: Cross-image relational knowledge distillation for semantic segmentation. In: CVPR (2022)
40. Ye, X., Dai, P., Luo, J., Guo, X., Qi, Y., Yang, J., Chen, Y.: Accelerating cnn training by pruning activation gradients (2020)
41. Yin, H., Molchanov, P., Alvarez, J.M., Li, Z., Mallya, A., Hoiem, D., Jha, N.K., Kautz, J.: Dreaming to distill: Data-free knowledge transfer via deepinversion. In: CVPR (2020)
42. Yoo, J., Cho, M., Kim, T., Kang, U.: Knowledge extraction with no observable data. In: NeurIPS (2019)

43. Yun, J., Lee, J., Kim, J., Kim, M., Song, J., Kang, U.: Sharvet: Similarity-aware parameter sharing with vector-based tuning for efficient llm compression. In: ACL (2026)
44. 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: CVPR (2022)