Novelty Assessment Report

Paper: LoRaQ: Optimized Low Rank Approximated Quantization Error for 4-bit Quantization

PDF URL: https://openreview.net/pdf?id=ECl6HGrMQI

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-30

Abstract

Post-training quantization (PTQ) is essential for deploying large diffusion-based transformers on resource-constrained hardware. However, aggressive 4-bit quantization introduces significant degradation in generative performance. While existing solutions mitigate quantization error through outlier smoothing or rotation techniques, low-rank approximation methods that add auxiliary linear branches to each quantized layer represent a promising new paradigm. Yet, these approaches suffer from computational overhead due to the data movement required by full-precision (W16A16) branches, limiting practical deployment. In addition, data-based calibration contributes to the computational complexity of the quantization process, especially because search policies must evaluate many parameter configurations using a small calibration subset. We propose LoRaQ (low-rank approximated quantization), a data-free calibration approach to optimize quantization error compensation. This method can be used in composition with other PTQ models. LoRaQ further enables mixed-precision configurations by quantizing the low-rank branch itself, overcoming the limitations of prior work. While LoRaQ achieves superior quantization performance than state-of-the-art methods in their native W4A4 setting on PixArt-Sigma and SANA, it also allows for configurations such as W8A8, W6A6 and W4A8 for low-rank branch alongside a W4 main layer. This reduces data movement overhead and enables a fully quantized, hardware-efficient solution.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Low-Rank Approximation for Post-Training Quantization of Diffusion Transformers**

A total of **46 papers** were analyzed and organized into a taxonomy with **19 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Quantization Techniques and Bit-Width Strategies**
- **Activation Outlier Management**
- **Temporal and Timestep-Aware Quantization**
- **Low-Rank Approximation and Decomposition**
- **Quantization-Aware Training and Fine-Tuning**
- **Post-Training Quantization Frameworks**
- **Unified Compression Strategies**
- **Specialized Quantization for Non-Image Modalities**
- **Surveys, Reviews, and Deployment Studies**
- **Specialized Architectures and Applications**

Complete Taxonomy Tree

- Low-Rank Approximation for Post-Training Quantization of Diffusion Transformers Survey Taxonomy
- Quantization Techniques and Bit-Width Strategies
    - Extreme Low-Bit Quantization (2-4 bit) ★ (5 papers)
    - [0] LoRaQ: Optimized Low Rank Approximated Quantization Error for 4-bit Quantization (Anon et al., 2026) View paper
    - [4] Mpq-dm: Mixed precision quantization for extremely low bit diffusion models (Feng Weilun, 2025) View paper
    - [8] Svdquant: Absorbing outliers by low-rank components for 4-bit diffusion models (Li, 2024) View paper
    - [9] Mpq-dmv2: Flexible residual mixed precision quantization for low-bit diffusion models with temporal distillation (Feng Weilun, 2025) View paper
    - [11] Terdit: Ternary diffusion models with transformers (Lu Xudong, 2024) View paper
    - Floating-Point and Hybrid Quantization (1 papers)
    - [1] Hq-dit: Efficient diffusion transformer with fp4 hybrid quantization (Liu, 2024) View paper
    - Standard and Mixed-Precision Quantization (2 papers)
    - [21] TreeQ: Pushing the Quantization Boundary of Diffusion Transformer via Tree-Structured Mixed-Precision Search (Kaicheng Yang, 2025) View paper
    - [27] MLoRQ: Bridging Low-Rank and Quantization for Transformer Compression (Gordon, 2025) View paper
    - Vector Quantization (1 papers)
    - [36] VQ4DiT: Efficient Post-Training Vector Quantization for Diffusion Transformers (Deng, 2025) View paper
- Activation Outlier Management
    - Smoothing and Transformation-Based Methods (3 papers)
    - [2] Ditas: Quantizing diffusion transformers via enhanced activation smoothing (Zhenyuan Dong, 2025) View paper
    - [22] HadaNorm: Diffusion Transformer Quantization through Mean-Centered Transformations (Federici, 2025) View paper
    - [39] HQ-DM: Single Hadamard Transformation-Based Quantization-Aware Training for Low-Bit Diffusion Models (Shizhuo Mao, 2025) View paper

  - Residual and Zero-Suppression Techniques (1 papers)
  - [26] Post-Training Quantization via Residual Truncation and Zero Suppression for Diffusion Models (Kim Dong-Hoon, 2025) View paper
- Temporal and Timestep-Aware Quantization
  - Timestep-Adaptive Quantization (3 papers)
  - [5] Passionsr: Post-training quantization with adaptive scale in one-step diffusion based image super-resolution (Libo Zhu, 2025) View paper
  - [14] Post-Training Quantization for Diffusion Transformer via Hierarchical Timestep Grouping (Ding Ning, 2025) View paper
  - [42] Towards Accurate Post-Training Quantization for Diffusion Models (Changyuan Wang, 2024) View paper
  - Time-Rotation Quantization (1 papers)
  - [3] Tr-dq: Time-rotation diffusion quantization (Shao Yihua, 2025) View paper
- Low-Rank Approximation and Decomposition
  - Low-Rank Branch for Outlier Absorption (1 papers)
  - [6] Post-Training Quantization for Audio Diffusion Transformers (Khandelwal, 2025) View paper
  - Unified Low-Rank and Quantization Frameworks (1 papers)
  - [28] UniQL: Unified Quantization and Low-rank Compression for Adaptive Edge LLMs (Hung-Yueh Chiang, 2025) View paper
- Quantization-Aware Training and Fine-Tuning
  - Selective Fine-Tuning for Quantization (1 papers)
  - [7] Quest: Low-bit diffusion model quantization via efficient selective finetuning (Wang, 2025) View paper
  - Full Quantization-Aware Training (1 papers)
  - [13] Efficientdm: Efficient quantization-aware fine-tuning of low-bit diffusion models (He, 2023) View paper
  - Fine-Tuning Quantized Models with Adapters (2 papers)
  - [18] Memory-efficient fine-tuning for quantized diffusion model (Hyogon Ryu, 2024) View paper
  - [35] IntLoRA: Integral Low-rank Adaptation of Quantized Diffusion Models (Guo Hang, 2025) View paper
- Post-Training Quantization Frameworks
  - General Post-Training Quantization for Diffusion Transformers (5 papers)
  - [23] PTQ4DiT: Post-training Quantization for Diffusion Transformers (Wu Junyi, 2024) View paper
  - [25] Q-DiT: Accurate Post-Training Quantization for Diffusion Transformers (Lei Chen, 2025) View paper
  - [29] Post-Training Quantization on Diffusion Models (Yuzhang Shang, 2023) View paper
  - [41] Q-Diffusion: Quantizing Diffusion Models (Xiuyu Li, 2023) View paper
  - [43] PTQD: Accurate Post-Training Quantization for Diffusion Models (He, 2023) View paper
  - Video and Audio Diffusion Quantization (3 papers)
  - [15] DVD-Quant: Data-free Video Diffusion Transformers Quantization (Li Zhiteng, 2025) View paper
  - [16] Q-VDiT: Towards Accurate Quantization and Distillation of Video-Generation Diffusion Transformers (Feng Weilun, 2025) View paper
  - [30] QVD: Post-training Quantization for Video Diffusion Models (Shilong Tian, 2024) View paper
- Unified Compression Strategies
  - Quantization with Attention Sparsification (1 papers)
  - [17] QuantSparse: Comprehensively Compressing Video Diffusion Transformer with Model Quantization and Attention Sparsification (Feng Weilun, 2025) View paper
- Specialized Quantization for Non-Image Modalities (2 papers)
  - [37] Quant-dLLM: Post-Training Extreme Low-Bit Quantization for Diffusion Large Language Models (Zhang Tianao, 2025) View paper
  - [46] Outlier-Aware Post-Training Quantization for Discrete Graph Diffusion Models (Z Gong, n.d.) View paper
- Surveys, Reviews, and Deployment Studies (10 papers)
  - [10] Diffusion Model Quantization: A Review (Zeng Qian, 2025) View paper
  - [12] Quantization as a Foundation for Deployable High Performance Diffusion Models within the Landscape of Large Scale Generative AI (Mikkel SÃ¸rensen, 2025) View paper
  - [24] Adaptive Compression and Quantization Techniques for Robust and Scalable Generative Diffusion Networks (Wei Li, 2025) View paper
  - [31] Low-Bit Generative Modeling with Diffusion Networks for Scalable and Perception-Aware Synthesis (Chand Aline, 2025) View paper
  - [32] Quantizing Diffusion Models for Scalable and Efficient Generative Inference Across Diverse Hardware Platforms (Markus Feldner, 2025) View paper
  - [33] From High Precision Denoising to Lightweight Generation with Quantized Diffusion Models (Chand Aline, 2025) View paper
  - [34] Strategies for Deploying High-Fidelity Generative Diffusion Models at Scale under Computational and Energy Constraints (Maria Jensen, 2025) View paper
  - [38] Techniques for Maintaining Stability and High-Fidelity Outputs in Resource-Constrained Deployments of Large Generative Models (L Petersen, 2025) View paper
  - [40] Towards Efficient Inference of Large Visual Generative Models (Dong, 2025) View paper
  - [44] Deliverable 02 (A Thirunavukkarasu-DLR, 2024) View paper
- Specialized Architectures and Applications (3 papers)
  - [19] An Analysis on Quantizing Diffusion Transformers (Yang Yue-wei, 2024) View paper
  - [20] MSDT: Multiscale Diffusion Transformer for Multimodality Image Fusion (Caifeng Xia, 2025) View paper
  - [45] VETA-DiT: Variance-Equalized and Temporally Adaptive Quantization for Efficient 4-bit Diffusion Transformers (L Yang, n.d.) View paper

Narrative

Core task: Low-rank approximation for post-training quantization of diffusion transformers. The field has organized itself around several complementary strategies for compressing diffusion models without extensive retraining. At the highest level, researchers pursue diverse bit-width strategies—ranging from moderate 8-bit schemes to extreme 2–4 bit quantization—while simultaneously addressing activation outliers that destabilize low-precision inference. A substantial body of work explores temporal and timestep-aware methods that adapt quantization parameters across the denoising trajectory, recognizing that different diffusion steps exhibit distinct numerical characteristics. Low-rank decomposition techniques offer an orthogonal avenue for reducing parameter counts, and many recent efforts combine quantization with such factorizations to achieve greater compression. Meanwhile, some branches focus on unified compression strategies that merge pruning, distillation, and quantization, or specialize in non-image modalities such as audio and video generation, reflecting the broadening scope of diffusion applications.

Within the extreme low-bit quantization branch, a particularly active line of work tackles the challenge of maintaining generation quality at 2–4 bits. Methods such as Mpq-dm[4], Svdquant[8], and Mpq-dmv2[9] demonstrate that careful calibration and mixed-precision allocation can preserve fidelity even under severe bit constraints, while Terdit[11] explores ternary representations. LoRaQ[0] situates itself in this cluster by integrating low-rank approximation directly into the post-training quantization pipeline, aiming to recover representational capacity lost during aggressive bit reduction. Compared to neighbors like Svdquant[8], which emphasizes singular value decomposition for weight matrices, LoRaQ[0] leverages rank-constrained updates to compensate for quantization error without requiring full fine-tuning. This approach contrasts with Mpq-dm[4] and Mpq-dmv2[9], which rely more heavily on mixed-precision search and calibration heuristics, highlighting an ongoing exploration of whether structural decomposition or adaptive precision offers a more scalable path to ultra-low-bit diffusion inference.

Related Works in Same Category

The following **3 sibling papers** share the same taxonomy leaf node with the original paper:

1. Mpq-dm: Mixed precision quantization for extremely low bit diffusion models

Authors: Feng Weilun, Wei Feng, Qin, Haotong, Haotong Qin, et al. (27 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Diffusion models have received wide attention in generation tasks. However, the expensive computation cost prevents the application of diffusion models in resource-constrained scenarios. Quantization emerges as a practical solution that significantly saves storage and computation by reducing the bit-width of parameters. However, the existing quantization methods for diffusion models still cause severe degradation in performance, especially under extremely low bit-widths (2-4 bit). The primary de...

Relationship Analysis

Both papers belong to the extreme low-bit quantization (2-4 bit) category for diffusion transformers, addressing the challenge of aggressive quantization through novel techniques. While LoRaQ focuses on data-free calibration using optimized low-rank approximation to compensate quantization error and enables mixed-precision configurations for the low-rank branch itself, MPQ-DM tackles the problem through outlier-driven mixed quantization within layers and time-smoothed relation distillation for robust optimization. The key distinction is that LoRaQ optimizes the low-rank matrices to absorb quantization error without data-dependent calibration, whereas MPQ-DM uses intra-layer mixed-precision based on Kurtosis to handle outlier channels and employs knowledge distillation across time steps.

2. Svdquant: Absorbing outliers by low-rank components for 4-bit diffusion models

Authors: Li, Muyang, Lin Yu-jun, Muyang Li, Zhang Zhekai, et al. (25 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Diffusion models can effectively generate high-quality images. However, as they scale, rising memory demands and higher latency pose substantial deployment challenges. In this work, we aim to accelerate diffusion models by quantizing their weights and activations to 4 bits. At such an aggressive level, both weights and activations are highly sensitive, where existing post-training quantization methods like smoothing become insufficient. To overcome this limitation, we propose SVDQuant, a new 4-b...

Relationship Analysis

Both papers belong to the extreme low-bit quantization category, focusing on 4-bit post-training quantization for diffusion transformers using low-rank decomposition techniques. They share the core approach of using low-rank branches to compensate for quantization error in aggressive 4-bit settings (W4A4), with both employing SVD-based decomposition and smoothing techniques. The key difference is that LoRaQ proposes a data-free calibration method that optimizes quantization error directly and quantizes the low-rank branch itself to various sub-16-bit formats (W4-W8 for low-rank branch), while SVDQuant maintains a full-precision (16-bit) low-rank branch and relies on data-dependent calibration with custom fused kernels for efficiency.

3. Terdit: Ternary diffusion models with transformers

Authors: Lu Xudong, Zhou Aojun, Xudong Lu, Lin Ziyi, Aojun Zhou, et al. (23 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Recent developments in large-scale pre-trained text-to-image diffusion models have significantly improved the generation of high-fidelity images, particularly with the emergence of diffusion transformer models (DiTs). Among diffusion models, diffusion transformers have demonstrated superior image-generation capabilities, boosting lower FID scores and higher scalability. However, deploying large-scale DiT models can be expensive due to their excessive parameter numbers. Although existing research...

Relationship Analysis

Both papers belong to the extreme low-bit quantization category (2-4 bit) for diffusion transformers, addressing the challenge of aggressive quantization while maintaining generative quality. While LoRaQ focuses on post-training quantization (PTQ) using low-rank approximation to compensate for quantization error in a data-free manner with mixed-precision configurations (W4A4, W4A8, etc.), TerDiT employs quantization-aware training (QAT) from scratch to achieve ternary (2-bit) weight quantization with full-precision activations, requiring complete model retraining rather than post-training calibration.

Contributions Analysis

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: LoRaQ: Data-free calibration approach for optimizing quantization error compensation

Description: The authors introduce LoRaQ, a post-training quantization method that optimizes quantization error by directly approximating the error using low-rank matrices through gradient descent, eliminating the need for data-dependent calibration. This approach can be composed with other PTQ models and simplifies the quantization process.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. AIQViT: Architecture-Informed Post-Training Quantization for Vision Transformers

URL: View paper

Brief Assessment

AIQViT[58] focuses on vision transformers with architecture-informed low-rank compensation using network architecture search to determine ranks, while LoRaQ targets diffusion transformers with data-free gradient-based optimization of low-rank matrices. The technical approaches and application domains differ substantially.

2. LQER: Low-Rank Quantization Error Reconstruction for LLMs

URL: View paper

Prior Art Analysis

LQER[63] demonstrates that similar prior work exists for data-free post-training quantization using low-rank approximation for error compensation. Both papers propose methods that approximate quantization error using low-rank matrices without requiring calibration datasets. LQER[63] explicitly states it requires 'no further weight training' and uses SVD-based low-rank approximation to reconstruct quantization error (eq = w - wq), which is conceptually identical to the original paper's approach of 'directly approximating the error using low-rank matrices through gradient descent.' The key technical overlap is that both methods decompose quantization error into low-rank components that can be optimized without data-dependent calibration, though they differ in optimization strategies (SVD vs. gradient descent).

Evidence

Evidence 1 - **Rationale**: Both papers emphasize post-training quantization without additional training/calibration, establishing the data-free constraint that is central to the original contribution claim. - **Original**: we propose loraq (low-rank approximated quantization), a data-free calibration approach to optimize quantization error compensation. this method can be used in composition with other ptq models. - **Candidate**: low-precision post-training quantization (ptq) of llms has recently become an attractive solution for reducing computational and memory cost (nagel et al., 2021). however, it remains challenging due to the fact that 1) no further weight training is allowed

Evidence 2 - **Rationale**: Both methods explicitly formulate the problem as reconstructing/approximating quantization error (eq = w - wq) using low-rank decomposition, demonstrating prior work on the core technical approach. - **Original**: we propose loraq (low-rank approximated quantization), a data-free calibration approach to optimize quantization error compensation... loraq further enables mixed-precision configurations by quantizing the low-rank branch itself - **Candidate**: our idea is to reconstruct the quantization error matrix eq through svd-based low rank approximation. when a quantization is applied to a trained fp32/fp16 weight matrix w ∈ rmxn, the resulting quantization error matrix eq is: eq = w - wq

Evidence 3 - **Rationale**: LQER[63] explicitly describes a framework combining quantization and low-rank approximation for error reduction, published prior to the original submission, refuting the novelty claim of being first to propose this approach. - **Original**: we propose a data-free optimization framework that directly minimizes quantization error via the low-rank branch, enabling more aggressive and efficient quantization of both weights and activations. - **Candidate**: we introduce a novel quantized llm inference framework termed low-rank quantization error reduction (lqer) which combines quantization and low-rank approximation. unlike existing methods that require gathering values from irregular memory locations, lqer boasts a blocked and regular computation patt...

Evidence 4 - **Rationale**: Both methods decompose the weight matrix into a quantized component (wq) plus a low-rank error correction term, showing LQER[63] proposed this dual-branch architecture before the original paper. - **Original**: loraq optimizes the low-rank matrices, which in turn defines the residual weights and enables the quantization of the low-rank matrix itself. - **Candidate**: if two high-precision matrices ak = uk and bk = σkv t k are assigned to approximate eq, i.e., akbk ≈ eq, the linear layer can be approximated as: ey = xwq + (xak)bk = x(wq + akbk) = x(wq + feq) ≈ x(wq + eq) = xw

Evidence 5 - **Rationale**: While the original paper uses gradient descent optimization, LQER[63] uses SVD decomposition to achieve the same goal of low-rank quantization error approximation, demonstrating prior work on the fundamental concept even with different optimization approaches. - **Original**: we propose to solve the optimization problem l∗, r∗ = arg min l,r ‖q(w + lr) - w - lr‖f ∝ arg min l,r l(q(w + lr) - w, lr) - **Candidate**: our idea is to reconstruct the quantization error matrix eq through svd-based low rank approximation... eq = uσv t ≈ ukσkv t k where u ∈ rmxm and v ∈ rnxn are orthogonal matrices, σ ∈ rmxn is a diagonal matrix of singular values.

3. Exploring Post-training Quantization in LLMs from Comprehensive Study to Low Rank Compensation

URL: View paper

Brief Assessment

Exploring Post-training Quantization in[59] focuses on LLMs with low-rank compensation (LoRC) applied after SVD decomposition of quantization error, while the original paper targets diffusion transformers with direct gradient-based optimization of low-rank matrices to absorb quantization error without data calibration.

4. ASER: activation smoothing and error reconstruction for large language model quantization

URL: View paper

Prior Art Analysis

ASER[57] demonstrates that similar prior work exists for using low-rank approximation to compensate quantization error without data-dependent calibration. Both papers propose using LoRA-style low-rank matrices to reconstruct quantization error through optimization. ASER[57] explicitly describes 'error reconstruction: low-rank compensation for quantization error with lora-style matrices' and states that 'lora-style matrices are used to reconstruct the quantization error, with little computation overhead.' The candidate paper's approach of identifying low-rank structure in quantization error and using gradient-based optimization to find compensating matrices directly parallels the original paper's core contribution. Both methods eliminate the need for calibration datasets and can be composed with other quantization techniques.

Evidence

Evidence 1 - **Rationale**: Both papers claim to propose a method that uses low-rank compensation for quantization error and can be composed with other quantization methods, demonstrating that ASER[57] presents similar prior work. - **Original**: we propose loraq (low-rank approximated quantization), a data-free calibration approach to optimize quantization error compensation. this method can be used in composition with other ptq models. - **Candidate**: we propose aser (activation smoothing and error reconstruction), a low-rank compensation algorithm designed to enhance the efficacy of quantized models... this lightweight yet powerful mechanism ensures robust performance of quantized models, which is orthogonal to any particular weight quantization...

Evidence 2 - **Rationale**: Both papers identify that quantization error has low-rank structure and propose using low-rank matrices specifically to compensate for this error rather than approximating the original weights, showing ASER[57] explored this concept prior to the original paper. - **Original**: we argue that this strategy is more effective than approximating the weight matrix itself with a low-rank decomposition, as our method provides explicit quantization error compensation, which adapts to the quantization operator and the chosen data format. - **Candidate**: we empirically find this inevitable quantization error has low-rank property, whose singular value distribution featuring a small number of high values and a long-tail bulk of low ones... which prompts us to consider using a module similar to lora (hu et al. 2021) to reconstruct this error.

Evidence 3 - **Rationale**: Both papers formulate the quantization problem as an optimization objective involving the difference between original and quantized weights, with ASER[57] showing prior work on this formulation approach. - **Original**: instead, we can reformulate the problem by considering the quantization function as an idempotent operator... thus, we propose to solve the optimization problem l∗, r∗ = arg min l,r ‖q(w + lr) - w - lr‖f - **Candidate**: for typical quantization, the optimization objective can be formulated as minimizing ‖wx - wqx‖f , where w and wq = q(w) are original model weight and its quantized one respectively.

5. Post-Training Quantization for Audio Diffusion Transformers

URL: View paper

Brief Assessment

Post-Training Quantization for Audio[6] focuses on audio diffusion transformers with timestep-aware smoothing and SVD-based LoRA branches, not on data-free calibration through gradient-based low-rank error optimization as proposed in the original paper.

6. Zeroquant-fp: A leap forward in llms post-training w4a8 quantization using floating-point formats

URL: View paper

Brief Assessment

Zeroquant-fp[55] focuses on floating-point quantization (FP8/FP4) for LLMs with weight-activation quantization schemes, while the original paper addresses diffusion transformers using low-rank approximation with data-free optimization. The technical domains and model architectures differ substantially.

7. RILQ: Rank-Insensitive LoRA-based Quantization Error Compensation for Boosting 2-bit Large Language Model Accuracy

URL: View paper

Brief Assessment

RILQ[56] focuses on rank-insensitive LoRA-based quantization error compensation for 2-bit LLMs using model-wise activation discrepancy loss, while the original paper addresses data-free calibration for 4-bit quantization in diffusion transformers using direct weight error approximation with low-rank matrices.

8. QSLR: Post-Training Compression via Quantized Sparse and Low-Rank Factorization

URL: View paper

Brief Assessment

QSLR[60] focuses on Hessian-aware quantization combined with sparse+low-rank decomposition, requiring calibration data for Hessian computation. The original paper's data-free gradient descent optimization for low-rank error compensation is technically distinct.

9. FIMA-Q: Post-Training Quantization for Vision Transformers by Fisher Information Matrix Approximation

URL: View paper

Brief Assessment

FIMA-Q[61] focuses on Fisher Information Matrix approximation for Vision Transformers, not on low-rank approximation methods for diffusion models. The technical approaches and application domains are fundamentally different.

10. ZeroQuant-V2: Exploring Post-training Quantization in LLMs from Comprehensive Study to Low Rank Compensation

URL: View paper

Brief Assessment

ZeroQuant-V2[62] focuses on comprehensive PTQ analysis and introduces LoRC (Low Rank Compensation) using SVD on quantization error matrices, while the original paper proposes LoRaQ with gradient descent optimization to directly approximate quantization error. The methods differ fundamentally in their optimization approaches and technical implementation.

Contribution 2: Mixed-precision quantization strategy with quantized low-rank branches

Description: The method enables flexible mixed-precision configurations (such as W8A8, W6A6, and W4A8) for the low-rank branch alongside a W4 main layer, reducing data movement overhead and enabling a fully quantized, hardware-efficient solution without requiring full-precision operations.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Neural precision polarization: Simplifying neural network inference with dual-level precision

URL: View paper

Prior Art Analysis

Neural precision polarization[71] demonstrates prior work on mixed-precision quantization with quantized low-rank branches. The candidate paper explicitly describes a dual-level precision approach where low-rank approximation paths are quantized to various precision levels (e.g., fp4, nf4, fp8, int8) alongside ultra-low precision main paths. This directly parallels the original paper's claim of enabling flexible mixed-precision configurations (W8A8, W6A6, W4A8) for low-rank branches alongside W4 main layers. Both papers address the same fundamental problem: reducing data movement overhead by quantizing low-rank branches rather than maintaining them at full precision, and both enable hardware-efficient solutions without requiring full-precision operations.

Evidence

Evidence 1 - **Rationale**: Both papers describe quantizing low-rank paths to enable edge deployment with reduced precision, addressing data movement overhead. - **Original**: we propose loraq (low-rank approximated quantization), a data-free calibration approach to optimize quantization error compensation. this method can be used in composition with other ptq models. loraq further enables mixed-precision configurations by quantizing the low-rank branch itself, overcoming... - **Candidate**: under npp, a cloud-trained floating-point model is downloaded to the edge, with weights and activations quantized to meet the edge. layer-wise surrogate paths with lowrank approximations and sensitivity-based metrics to recover accuracy with minimal overhead and retraining data. computein-memory pro...

Evidence 2 - **Rationale**: Both papers explicitly contrast their approach with maintaining full-precision low-rank branches, proposing quantized low-rank branches instead for efficiency. - **Original**: unlike li et al. (2025)'s approach of maintaining a full-precision low-rank branch, we propose a quantized low-rank branch to reduce data movement overhead, eliminating custom kernel requirements and enabling flexible mixed-precision weight matrices. - **Candidate**: in contrast, our neural precision polarization (npp) dedicates the main processing path to ultra-low precision (e.g., 4-bit or 8-bit), while high precision is reserved solely for surrogate quantization error and process variability compensation paths. by clearly distinguishing precision requirements...

Evidence 3 - **Rationale**: Both papers describe quantizing low-rank branches at various precision levels (fp4, nf4, fp8, etc.) to balance accuracy recovery with computational efficiency. - **Original**: we can define a budget β as the maximum allowable memory and computation resources for the low-rank branch in our approximation. β is defined by the number of bits we require per channel for the low-rank branch. for example, svdquant uses a float16 low-rank branch with ρ = 32 for its 4-bit quantizat... - **Candidate**: figure 2 shows that accuracy loss from ultra-low-precision processing paths can be effectively mitigated with lorabased high-precision surrogate paths. this approach, tested on transformer based models, specifically vit with cifar100 and imagenet, showed that extreme quantization (e.g., fp4, nf4) ca...

Evidence 4 - **Rationale**: Both papers describe dual quantization functions/levels: one for the main/residual branch at ultra-low precision and another for the low-rank/surrogate branch at sub-16-bit precision. - **Original**: we find that the accuracy gains from using a higher rank outweigh the loss from quantizing the low-rank matrices. reminding the description of the method in figure 1, we now consider two quantization functions q1 for the residual branch and q2 for the low-rank branch to be represented in n <16 bits/... - **Candidate**: we propose neural precision polarization (npp) , implementing processing with only two precision levels. in fig. 1, under npp the majority of network weights are processed in ultra-low precision (e.g., nf4 or fp4) to meet device constraints, with limited high-precision paths added to recover accurac...

2. Delta-come: Training-free delta-compression with mixed-precision for large language models

URL: View paper

Brief Assessment

Delta-come[68] focuses on delta-compression for fine-tuned LLMs using mixed-precision quantization of singular vectors based on singular value magnitudes. The original paper addresses post-training quantization for diffusion transformers with quantized low-rank branches for error compensation. These are fundamentally different application domains and technical approaches.

3. Adaptive quantization error reconstruction for llms with mixed precision

URL: View paper

Brief Assessment

Adaptive quantization error reconstruction[64] focuses on adaptive mixed-precision quantization for LLMs with low-rank error reconstruction based on output activations, not on mixed-precision configurations for diffusion transformers with quantized low-rank branches alongside residual layers as in the original paper.

4. LoRAQuant: Mixed-Precision Quantization of LoRA to Ultra-Low Bits

URL: View paper

Brief Assessment

LoRAQuant[72] focuses on quantizing LoRA adapters for LLMs using SVD-based reparameterization, while the original paper addresses quantization of diffusion transformer weights with low-rank branches for image generation tasks. These are fundamentally different application domains and architectural contexts.

5. Efficient Fine-Tuning of Quantized Models via Adaptive Rank and Bitwidth

URL: View paper

Brief Assessment

Efficient Fine-Tuning of Quantized[66] focuses on adaptive rank and bitwidth selection for fine-tuning quantized models using calibration data, while the original paper addresses post-training quantization for inference with data-free optimization of low-rank branches.

6. Collaborative automotive radar sensing via mixed-precision distributed array completion

URL: View paper

Brief Assessment

Collaborative automotive radar sensing[69] addresses mixed-precision quantization in radar signal processing for automotive applications, not neural network quantization. The technical domains and problem formulations are fundamentally different.

7. MP-DPD: Low-Complexity Mixed-Precision Neural Networks for Energy-Efficient Digital Predistortion of Wideband Power Amplifiers

URL: View paper

Brief Assessment

MP-DPD[67] focuses on mixed-precision neural networks for digital predistortion in power amplifiers, not on low-rank branch quantization in diffusion models or general neural network architectures.

8. ResQ: Mixed-Precision Quantization of Large Language Models with Low-Rank Residuals

URL: View paper

Brief Assessment

ResQ[65] focuses on mixed-precision quantization for LLMs using PCA-based subspace identification, while the original paper targets diffusion transformers with SVD-based low-rank decomposition. The technical approaches and application domains differ fundamentally.

9. Quantformer: Learning extremely low-precision vision transformers

URL: View paper

Brief Assessment

Quantformer[70] focuses on vision transformers with self-attention rank preservation and group-wise patch feature quantization, not on low-rank branch decomposition for diffusion transformers. The mixed-precision integration mentioned is fundamentally different from the original paper's low-rank approximation approach.

Contribution 3: Open-source hardware-agnostic PTQ library for transformer blocks

Description: The authors provide an open-source library that enables systematic benchmarking of post-training quantization methods across different configurations and supports scalable quantization of large models in multi-GPU environments, facilitating reproducible research and practical deployment.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Efficient post-training quantization with fp8 formats

URL: View paper

Brief Assessment

Efficient post-training quantization with[53] focuses on FP8 quantization formats and workflows for various neural network architectures, not on providing an open-source PTQ library specifically for transformer blocks. The papers address different aspects of quantization.

2. ZeroQuant: Efficient and Affordable Post-Training Quantization for Large-Scale Transformers

URL: View paper

Brief Assessment

ZeroQuant[47] focuses on post-training quantization for transformer models with system optimization for inference speedup, but does not explicitly describe releasing an open-source library for systematic benchmarking across different PTQ methods and configurations as claimed in the original paper.

3. UniQL: Unified Quantization and Low-rank Compression for Adaptive Edge LLMs

URL: View paper

Brief Assessment

UniQL[28] focuses on unified quantization and low-rank compression for edge LLMs with on-device adaptive pruning, not on providing a general-purpose PTQ library for systematic benchmarking across different configurations as described in the original contribution.

4. Framequant: Flexible low-bit quantization for transformers

URL: View paper

Brief Assessment

Framequant[54] focuses on fusion frame-based quantization theory and does not describe an open-source library for systematic benchmarking or multi-GPU scalable quantization of transformer blocks.

5. Efficientqat: Efficient quantization-aware training for large language models

URL: View paper

Brief Assessment

Efficientqat[51] focuses on quantization-aware training (QAT) for LLMs, not post-training quantization (PTQ) libraries. The original paper provides a PTQ library for transformer blocks in diffusion models, while the candidate addresses QAT methodology for language models.

6. Q-DiT: Accurate Post-Training Quantization for Diffusion Transformers

URL: View paper

Brief Assessment

Q-DiT[25] focuses on quantization techniques for diffusion transformers (automatic granularity allocation and dynamic activation quantization) but does not describe an open-source library infrastructure for systematic benchmarking or multi-GPU scalable quantization.

7. Repquant: Towards accurate post-training quantization of large transformer models via scale reparameterization

URL: View paper

Brief Assessment

Repquant[50] focuses on scale reparameterization techniques for quantizing specific transformer components (LayerNorm, Softmax) rather than providing a general-purpose PTQ library infrastructure. The paper does not describe library features like systematic benchmarking tools, multi-GPU support, or reproducible research frameworks.

8. Repq-vit: Scale reparameterization for post-training quantization of vision transformers

URL: View paper

Brief Assessment

Repq-vit[49] focuses on vision transformers for image tasks with scale reparameterization techniques, not on providing a general PTQ library for transformer blocks across different modalities or supporting multi-GPU quantization of large models.

9. Gptq: Accurate post-training quantization for generative pre-trained transformers

URL: View paper

Brief Assessment

Gptq[48] focuses on a specific quantization algorithm (GPTQ) for generative pre-trained transformers, not on providing a general-purpose open-source library for systematic benchmarking of PTQ methods across different configurations and multi-GPU scalable quantization as described in the original contribution.

10. Sparsegpt: Massive language models can be accurately pruned in one-shot

URL: View paper

Brief Assessment

Sparsegpt[52] focuses on pruning (removing weights) rather than quantization (reducing precision). The candidate does not provide an open-source PTQ library for transformer blocks.

Appendix: Text Similarity Detection

Textual similarity detection checked 33 papers and found 2 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Svdquant: Absorbing outliers by low-rank components for 4-bit diffusion models

Detected in: Core Task (sibling)

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] LoRaQ: Optimized Low Rank Approximated Quantization Error for 4-bit Quantization View paper
- [1] Hq-dit: Efficient diffusion transformer with fp4 hybrid quantization View paper
- [2] Ditas: Quantizing diffusion transformers via enhanced activation smoothing View paper
- [3] Tr-dq: Time-rotation diffusion quantization View paper
- [4] Mpq-dm: Mixed precision quantization for extremely low bit diffusion models View paper
- [5] Passionsr: Post-training quantization with adaptive scale in one-step diffusion based image super-resolution View paper
- [6] Post-Training Quantization for Audio Diffusion Transformers View paper
- [7] Quest: Low-bit diffusion model quantization via efficient selective finetuning View paper

- [8] Svdquant: Absorbing outliers by low-rank components for 4-bit diffusion models View paper
- [9] Mpq-dmv2: Flexible residual mixed precision quantization for low-bit diffusion models with temporal distillation View paper
- [10] Diffusion Model Quantization: A Review View paper
- [11] Terdit: Ternary diffusion models with transformers View paper
- [12] Quantization as a Foundation for Deployable High Performance Diffusion Models within the Landscape of Large Scale Generative AI View paper
- [13] Efficientdm: Efficient quantization-aware fine-tuning of low-bit diffusion models View paper
- [14] Post-Training Quantization for Diffusion Transformer via Hierarchical Timestep Grouping View paper
- [15] DVD-Quant: Data-free Video Diffusion Transformers Quantization View paper
- [16] Q-VDiT: Towards Accurate Quantization and Distillation of Video-Generation Diffusion Transformers View paper
- [17] QuantSparse: Comprehensively Compressing Video Diffusion Transformer with Model Quantization and Attention Sparsification View paper
- [18] Memory-efficient fine-tuning for quantized diffusion model View paper
- [19] An Analysis on Quantizing Diffusion Transformers View paper
- [20] MSDT: Multiscale Diffusion Transformer for Multimodality Image Fusion View paper
- [21] TreeQ: Pushing the Quantization Boundary of Diffusion Transformer via Tree-Structured Mixed-Precision Search View paper
- [22] HadaNorm: Diffusion Transformer Quantization through Mean-Centered Transformations View paper
- [23] PTQ4DiT: Post-training Quantization for Diffusion Transformers View paper
- [24] Adaptive Compression and Quantization Techniques for Robust and Scalable Generative Diffusion Networks View paper
- [25] Q-DiT: Accurate Post-Training Quantization for Diffusion Transformers View paper
- [26] Post-Training Quantization via Residual Truncation and Zero Suppression for Diffusion Models View paper
- [27] MLoRQ: Bridging Low-Rank and Quantization for Transformer Compression View paper
- [28] UniQL: Unified Quantization and Low-rank Compression for Adaptive Edge LLMs View paper
- [29] Post-Training Quantization on Diffusion Models View paper
- [30] QVD: Post-training Quantization for Video Diffusion Models View paper
- [31] Low-Bit Generative Modeling with Diffusion Networks for Scalable and Perception-Aware Synthesis View paper
- [32] Quantizing Diffusion Models for Scalable and Efficient Generative Inference Across Diverse Hardware Platforms View paper
- [33] From High Precision Denoising to Lightweight Generation with Quantized Diffusion Models View paper
- [34] Strategies for Deploying High-Fidelity Generative Diffusion Models at Scale under Computational and Energy Constraints View paper
- [35] IntLoRA: Integral Low-rank Adaptation of Quantized Diffusion Models View paper
- [36] VQ4DiT: Efficient Post-Training Vector Quantization for Diffusion Transformers View paper
- [37] Quant-dLLM: Post-Training Extreme Low-Bit Quantization for Diffusion Large Language Models View paper
- [38] Techniques for Maintaining Stability and High-Fidelity Outputs in Resource-Constrained Deployments of Large Generative Models View paper
- [39] HQ-DM: Single Hadamard Transformation-Based Quantization-Aware Training for Low-Bit Diffusion Models View paper
- [40] Towards Efficient Inference of Large Visual Generative Models View paper
- [41] Q-Diffusion: Quantizing Diffusion Models View paper
- [42] Towards Accurate Post-Training Quantization for Diffusion Models View paper
- [43] PTQD: Accurate Post-Training Quantization for Diffusion Models View paper
- [44] Deliverable 02 View paper
- [45] VETA-DiT: Variance-Equalized and Temporally Adaptive Quantization for Efficient 4-bit Diffusion Transformers View paper
- [46] Outlier-Aware Post-Training Quantization for Discrete Graph Diffusion Models View paper
- [47] ZeroQuant: Efficient and Affordable Post-Training Quantization for Large-Scale Transformers View paper
- [48] Gptq: Accurate post-training quantization for generative pre-trained transformers View paper
- [49] Repq-vit: Scale reparameterization for post-training quantization of vision transformers View paper
- [50] Repquant: Towards accurate post-training quantization of large transformer models via scale reparameterization View paper
- [51] Efficientqat: Efficient quantization-aware training for large language models View paper
- [52] Sparsegpt: Massive language models can be accurately pruned in one-shot View paper
- [53] Efficient post-training quantization with fp8 formats View paper
- [54] Framequant: Flexible low-bit quantization for transformers View paper
- [55] Zeroquant-fp: A leap forward in llms post-training w4a8 quantization using floating-point formats View paper
- [56] RILQ: Rank-Insensitive LoRA-based Quantization Error Compensation for Boosting 2-bit Large Language Model Accuracy View paper
- [57] ASER: activation smoothing and error reconstruction for large language model quantization View paper
- [58] AIQViT: Architecture-Informed Post-Training Quantization for Vision Transformers View paper
- [59] Exploring Post-training Quantization in LLMs from Comprehensive Study to Low Rank Compensation View paper
- [60] QSLR: Post-Training Compression via Quantized Sparse and Low-Rank Factorization View paper
- [61] FIMA-Q: Post-Training Quantization for Vision Transformers by Fisher Information Matrix Approximation View paper
- [62] ZeroQuant-V2: Exploring Post-training Quantization in LLMs from Comprehensive Study to Low Rank Compensation View paper
- [63] LQER: Low-Rank Quantization Error Reconstruction for LLMs View paper
- [64] Adaptive quantization error reconstruction for llms with mixed precision View paper
- [65] ResQ: Mixed-Precision Quantization of Large Language Models with Low-Rank Residuals View paper
- [66] Efficient Fine-Tuning of Quantized Models via Adaptive Rank and Bitwidth View paper
- [67] MP-DPD: Low-Complexity Mixed-Precision Neural Networks for Energy-Efficient Digital Predistortion of Wideband Power Amplifiers View paper
- [68] Delta-come: Training-free delta-compression with mixed-precision for large language models View paper
- [69] Collaborative automotive radar sensing via mixed-precision distributed array completion View paper
- [70] Quantformer: Learning extremely low-precision vision transformers View paper
- [71] Neural precision polarization: Simplifying neural network inference with dual-level precision View paper
- [72] LoRAQuant: Mixed-Precision Quantization of LoRA to Ultra-Low Bits View paper