Novelty Assessment Report

Paper: QVGen: Pushing the Limit of Quantized Video Generative Models

PDF URL: https://openreview.net/pdf?id=XJXZXuTj11

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-29

Abstract

Video diffusion models (DMs) have enabled high-quality video synthesis. Yet, their substantial computational and memory demands pose serious challenges to real-world deployment, even on high-end GPUs. As a commonly adopted solution, quantization has proven notable success in reducing cost for image DMs, while its direct application to video DMs remains ineffective. In this paper, we present QVGen, a novel quantization-aware training (QAT) framework tailored for high-performance and inference-efficient video DMs under extremely low-bit quantization (e.g., 4-bit or below). We begin with a theoretical analysis demonstrating that reducing the gradient norm is essential to facilitate convergence for QAT. To this end, we introduce auxiliary modules (Φ) to mitigate large quantization errors, leading to significantly enhanced convergence. To eliminate the inference overhead of Φ, we propose a rank-decay strategy that progressively eliminates Φ. Specifically, we repeatedly employ singular value decomposition (SVD) and a proposed rank-based regularization $\mathbf{\gamma}$ to identify and decay low-contributing components. This strategy retains performance while zeroing out additional inference overhead. Extensive experiments across 4 state-of-the-art (SOTA) video DMs, with parameter sizes ranging from $1.3\text{B}\sim14\text{B}$, show that QVGen is the first to reach full-precision comparable quality under 4-bit settings. Moreover, it significantly outperforms existing methods. For instance, our 3-bit CogVideoX-2B achieves improvements of $+25.28$ in Dynamic Degree and $+8.43$ in Scene Consistency on VBench. Code and videos are available in the supplementary material.

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: **Quantization-Aware Training for Video Diffusion Models**

A total of **38 papers** were analyzed and organized into a taxonomy with **12 categories**.

Taxonomy Overview

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

- **Quantization Strategy and Optimization**
- **Feature-Aware Quantization**
- **Joint Optimization with Complementary Techniques**
- **Deployment-Oriented Quantization**
- **Theoretical Foundations and Comprehensive Surveys**
- **Application-Specific Quantization**

Complete Taxonomy Tree

- Quantization-Aware Training for Video Diffusion Models Survey Taxonomy
- Quantization Strategy and Optimization
  - Post-Training Quantization (PTQ) (7 papers)
  - [7] ViDiT-Q: Efficient and Accurate Quantization of Diffusion Transformers for Image and Video Generation (Zhao, 2024) View paper
  - [14] Hardware-Friendly Static Quantization Method for Video Diffusion Transformers (Sanghyun Yi, 2025) View paper
  - [17] Q-DiT: Accurate Post-Training Quantization for Diffusion Transformers (Lei Chen, 2024) View paper
  - [19] QVD: Post-training Quantization for Video Diffusion Models (Shilong Tian, 2024) View paper
  - [21] DVD-Quant: Data-free Video Diffusion Transformers Quantization (Li Zhiteng, 2025) View paper
  - [22] LRQ-DiT: Log-Rotation Post-Training Quantization of Diffusion Transformers for Image and Video Generation (Yang, 2025) View paper
  - [30] CLQ: Cross-Layer Guided Orthogonal-based Quantization for Diffusion Transformers (Liu Kai, 2025) View paper
  - Quantization-Aware Training (QAT) ★ (4 papers)
  - [0] QVGen: Pushing the Limit of Quantized Video Generative Models (Anon et al., 2026) View paper
  - [23] FraQAT: Quantization Aware Training with Fractional bits (Morreale, 2025) View paper
  - [31] DilateQuant: Accurate and Efficient Quantization-Aware Training for Diffusion Models via Weight Dilation (Xuewen Liu, 2025) View paper
  - [36] DilateQuant: Accurate and Efficient Diffusion Quantization via Weight Dilation (Liu Xuewen, 2024) View paper
  - Mixed-Precision and Dynamic Quantization (2 papers)
  - [28] Temporal Dynamic Quantization for Diffusion Models (So, 2023) View paper
  - [34] Dynamic and Mixed-Precision Techniques for Scalable Iterative Generative Modeling (Mikkel Jensen, 2025) View paper
- Feature-Aware Quantization
  - Temporal and Timestep-Adaptive Quantization (5 papers)
  - [1] Tr-dq: Time-rotation diffusion quantization (Shao Yihua, 2025) View paper
  - [2] TCAQ-DM: timestep-channel adaptive quantization for diffusion models (Huang Haocheng, 2025) View paper
  - [10] TaQ-DiT: Time-aware Quantization for Diffusion Transformers (Liu Xin-yan, 2024) View paper
  - [11] TFMQ-DM: Temporal Feature Maintenance Quantization for Diffusion Models (Yushi Huang, 2023) View paper

    - [38] VETA-DiT: Variance-Equalized and Temporally Adaptive Quantization for Efficient 4-bit Diffusion Transformers (L Yang, n.d.) View paper
    - Channel and Activation-Aware Quantization (1 papers)
    - [33] Quantizing Diffusion Models from a Sampling-Aware Perspective (Zeng Qian, 2025) View paper
- Joint Optimization with Complementary Techniques
    - Quantization with Sparsity Co-Design (2 papers)
    - [6] FPSAttention: Training-Aware FP8 and Sparsity Co-Design for Fast Video Diffusion (Liu, 2025) View paper
    - [18] QuantSparse: Comprehensively Compressing Video Diffusion Transformer with Model Quantization and Attention Sparsification (Feng Weilun, 2025) View paper
    - Quantization with Distillation or Caching (3 papers)
    - [4] Q-VDiT: Towards Accurate Quantization and Distillation of Video-Generation Diffusion Transformers (Feng Weilun, 2025) View paper
    - [9] QuantCache: Adaptive Importance-Guided Quantization with Hierarchical Latent and Layer Caching for Video Generation (Wu Junyi, 2025) View paper
    - [20] S^{2}Q-VDiT: Accurate Quantized Video Diffusion Transformer with Salient Data and Sparse Token Distillation (Feng Weilun, 2025) View paper
- Deployment-Oriented Quantization
    - Hardware-Aware and Edge Deployment (3 papers)
    - [3] PARO: Hardware-Software Co-design with Pattern-aware Reorder-based Attention Quantization in Video Generation Models (Xinhao Yang, 2025) View paper
    - [5] Quantization as a Foundation for Deployable High Performance Diffusion Models within the Landscape of Large Scale Generative AI (Mikkel SÃ¸rensen, 2025) View paper
    - [8] Securing Federated Diffusion Model With Dynamic Quantization for Generative AI Services in Multiple-Access Artificial Intelligence of Things (Jiayi He, 2024) View paper
    - Real-Time and Low-Latency Applications (1 papers)
    - [13] LLIA - Enabling Low-Latency Interactive Avatars: Real-Time Audio-Driven Portrait Video Generation with Diffusion Models (Yu, 2025) View paper
- Theoretical Foundations and Comprehensive Surveys (5 papers)
    - [24] Adaptive Compression and Quantization Techniques for Robust and Scalable Generative Diffusion Networks (Wei Li, 2025) View paper
    - [26] Low-Bit Generative Modeling with Diffusion Networks for Scalable and Perception-Aware Synthesis (Chand Aline, 2025) View paper
    - [27] Quantizing Diffusion Models for Scalable and Efficient Generative Inference Across Diverse Hardware Platforms (Markus Feldner, 2025) View paper
    - [29] From High Precision Denoising to Lightweight Generation with Quantized Diffusion Models (Chand Aline, 2025) View paper
    - [32] Contemporary Advances in Neural Network Quantization: A Survey (Min Li, 2024) View paper
- Application-Specific Quantization
    - Interactive and Controllable Video Generation (2 papers)
    - [12] Yume: An Interactive World Generation Model (Mao, 2025) View paper
    - [16] ARLON: Boosting Diffusion Transformers with Autoregressive Models for Long Video Generation (Li, 2024) View paper
    - Specialized Domain Applications (4 papers)
    - [15] Bitrate-Controlled Diffusion for Disentangling Motion and Content in Video (Li Xiao, 2025) View paper
    - [25] Advanced Sign Language Video Generation with Compressed and Quantized Multi-Condition Tokenization (Wang Cong, 2025) View paper
    - [35] UniHM: Universal Human Motion Generation with Object Interactions in Indoor Scenes (Hayder, 2025) View paper
    - [37] GFix: Perceptually Enhanced Gaussian Splatting Video Compression (Siyue Teng, 2025) View paper

Narrative

Core task: quantization-aware training for video diffusion models. The field has organized itself around several complementary perspectives on reducing the computational and memory footprint of diffusion-based video generation. At the highest level, Quantization Strategy and Optimization explores fundamental training regimes—including quantization-aware training (QAT), post-training quantization (PTQ), and mixed-precision schemes—that directly address how to learn or calibrate low-bit representations. Feature-Aware Quantization focuses on exploiting structural properties of activations, weights, or temporal dynamics to assign bits more intelligently across layers or time steps. Joint Optimization with Complementary Techniques investigates hybrid approaches that combine quantization with pruning, knowledge distillation, or low-rank decomposition to achieve greater compression. Deployment-Oriented Quantization emphasizes hardware constraints and real-world inference scenarios, while Theoretical Foundations and Comprehensive Surveys provide broader context on convergence guarantees and design principles. Finally, Application-Specific Quantization tailors methods to particular domains such as sign language or medical imaging, where domain priors can guide bit allocation.

Within the Quantization Strategy and Optimization branch, a dense cluster of works explores QAT variants that retrain or fine-tune diffusion models end-to-end with quantized operations. QVGen[0] exemplifies this direction by integrating quantization directly into the video diffusion training loop, aiming to preserve generation quality under aggressive bit-width reduction. Nearby efforts such as FraQAT[23] and DilateQuant[31] similarly adopt QAT but introduce specialized techniques—fractional bit allocations or dilated convolution-aware quantizers—to handle the unique temporal coherence demands of video. In contrast, methods like TCAQ[2] and Time-Rotation Diffusion Quantization[1] emphasize calibration strategies that adapt quantization parameters across diffusion timesteps or rotational embeddings, blurring the line between pure QAT and hybrid calibration. The central trade-off across these lines is whether to invest training compute for tighter integration (as QVGen[0] does) or to rely on lighter post-hoc adjustments that may sacrifice some quality but reduce retraining overhead.

Related Works in Same Category

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

1. FraQAT: Quantization Aware Training with Fractional bits

Authors: Morreale, Luca, Ramos, Alberto Gil C. P., Luca Morreale, et al. (19 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

State-of-the-art (SOTA) generative models have demonstrated impressive capabilities in image synthesis or text generation, often with a large capacity model. However, these large models cannot be deployed on smartphones due to the limited availability of on-board memory and computations. Quantization methods lower the precision of the model parameters, allowing for efficient computations, \eg, in \INT{8}. Although aggressive quantization addresses efficiency and memory constraints, preserving th...

Relationship Analysis

Both papers belong to the Quantization-Aware Training (QAT) category, integrating quantization into the training process with gradient-based optimization to minimize quantization-induced degradation. They overlap in addressing low-bit quantization (4-bit and below) for diffusion models through training-based approaches. However, the original paper (QVGen) focuses specifically on video diffusion models with auxiliary modules and rank-decay strategies to improve convergence, while the candidate paper (FraQAT) targets text-to-image diffusion models using a fractional bits curriculum learning approach that progressively reduces precision from 32 to 4 bits during training.

2. DilateQuant: Accurate and Efficient Quantization-Aware Training for Diffusion Models via Weight Dilation

Authors: Xuewen Liu, Zhikai Li, Minghao Jiang, Mengjuan Chen, Minhao Jiang, et al. (7 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Model quantization is a promising method for accelerating and compressing diffusion models. Nevertheless, since post-training quantization (PTQ) fails catastrophically at low-bit cases, quantization-aware training (QAT) is essential. Unfortunately, the wide range and time-varying activations in diffusion models sharply increase the complexity of quantization, making existing QAT methods inefficient. Equivalent scaling can effectively reduce activation range, but previous methods remain the overa...

Relationship Analysis

Both papers belong to the Quantization-Aware Training (QAT) category, integrating quantization into the training process with gradient-based optimization for video diffusion models. They overlap in addressing convergence challenges during low-bit QAT through auxiliary mechanisms—QVGen introduces auxiliary modules Φ with a rank-decay strategy to reduce gradient norms, while DilateQuant proposes Weight Dilation to preserve weight distributions and uses Temporal Parallel Quantizer for time-varying activations. The key difference is that QVGen focuses on progressively eliminating auxiliary modules via SVD-based rank decay to avoid inference overhead, whereas DilateQuant emphasizes equivalent scaling techniques that maintain original weight ranges and employs block-wise knowledge distillation for efficiency.

3. DilateQuant: Accurate and Efficient Diffusion Quantization via Weight Dilation

Authors: Liu Xuewen, Xuewen Liu, Li, Zhikai, Zhikai Li, et al. (12 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Model quantization is a promising method for accelerating and compressing diffusion models. Nevertheless, since post-training quantization (PTQ) fails catastrophically at low-bit cases, quantization-aware training (QAT) is essential. Unfortunately, the wide range and time-varying activations in diffusion models sharply increase the complexity of quantization, making existing QAT methods inefficient. Equivalent scaling can effectively reduce activation range, but previous methods remain the overa...

Relationship Analysis

Both papers belong to the Quantization-Aware Training (QAT) category, integrating quantization into the training process with gradient-based optimization. They overlap in addressing quantization challenges for diffusion models, including handling wide activation ranges and achieving low-bit quantization (3-4 bits). However, the original paper (QVGen) focuses specifically on video diffusion models with auxiliary modules and rank-decay strategies, while the candidate paper (DilateQuant) targets general diffusion models (primarily image generation) using weight dilation and temporal parallel quantization techniques.

Contributions Analysis

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

Contribution 1: QVGen: A novel QAT framework for video diffusion models

Description: The authors present QVGen, the first quantization-aware training framework specifically designed for video diffusion models. It enables effective 3-bit and 4-bit quantization while achieving full-precision comparable quality, addressing the challenge that existing QAT methods fail to handle video generation tasks under extremely low-bit settings.

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. MPQ-DM: Mixed Precision Quantization for Extremely Low Bit Diffusion Models

URL: View paper

Brief Assessment

MPQ-DM[57] focuses on mixed-precision quantization for diffusion models without specifically targeting video generation tasks or QAT frameworks for video diffusion models.

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

URL: View paper

Brief Assessment

Q-DiT[17] focuses on post-training quantization (PTQ) for diffusion transformers, not quantization-aware training (QAT). The original paper explicitly presents QVGen as 'the first quantization-aware training framework specifically designed for video diffusion models,' while Q-DiT[17] is a PTQ method that does not involve retraining.

3. Quantization as a Foundation for Deployable High Performance Diffusion Models within the Landscape of Large Scale Generative AI

URL: View paper

Brief Assessment

Quantization Foundation Diffusion[5] discusses quantization techniques broadly for diffusion models but does not present a specific QAT framework for video diffusion models that would challenge QVGen's novelty claim as the first such framework.

4. MPQ-DMv2: Flexible Residual Mixed Precision Quantization for Low-Bit Diffusion Models with Temporal Distillation

URL: View paper

Brief Assessment

MPQ-DMv2[58] focuses on mixed-precision quantization with residual branches and temporal distillation for diffusion models, not specifically on video generation QAT frameworks or gradient norm reduction strategies.

5. PTQ4DiT: Post-training Quantization for Diffusion Transformers

URL: View paper

Brief Assessment

PTQ4DiT[55] focuses on post-training quantization for diffusion transformers in image generation, not quantization-aware training for video diffusion models. The technical approaches differ fundamentally (PTQ vs. QAT).

6. ViDiT-Q: Efficient and Accurate Quantization of Diffusion Transformers for Image and Video Generation

URL: View paper

Brief Assessment

ViDiT-Q[7] focuses on post-training quantization (PTQ) for diffusion transformers, not quantization-aware training (QAT). The candidate explicitly states it uses PTQ methods, which is fundamentally different from QVGen's QAT approach.

7. Q-dm: An efficient low-bit quantized diffusion model

URL: View paper

Brief Assessment

Q-DM[56] focuses on image diffusion models (DDPM/DDIM on CIFAR-10 and ImageNet), not video diffusion models. The candidate addresses different technical challenges (activation distribution oscillation across timesteps, quantization error accumulation) but does not demonstrate prior work on quantization-aware training specifically for video generation tasks.

8. QVD: Post-training Quantization for Video Diffusion Models

URL: View paper

Brief Assessment

QVD[19] focuses on post-training quantization (PTQ) for video diffusion models, not quantization-aware training (QAT). The paper explicitly states it is 'the first PTQ method tailored explicitly for video diffusion models' and does not involve retraining or fine-tuning, which is fundamentally different from QAT approaches.

9. Bidm: Pushing the limit of quantization for diffusion models

URL: View paper

Brief Assessment

BiDM[54] focuses on fully binarizing (1-bit) image diffusion models, not video diffusion models. The candidate addresses extreme quantization for image generation tasks using pixel-space and latent-space diffusion models, whereas the original contribution specifically targets video diffusion models with 3-bit and 4-bit quantization.

Contribution 2: Auxiliary modules (Φ) to reduce gradient norm and improve convergence

Description: The authors introduce learnable auxiliary modules that mitigate quantization errors during training. Through theoretical analysis demonstrating that reducing gradient norm is essential for QAT convergence, these modules stabilize the training process and significantly enhance convergence for extremely low-bit quantization.

This contribution was assessed against **5 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Stable Quantization-Aware Training with Adaptive Gradient Clipping

URL: View paper

Brief Assessment

Stable QAT Clipping[52] focuses on adaptive gradient clipping and dropout for stable QAT, not on auxiliary modules that mitigate quantization errors to reduce gradient norm.

2. Training Quantized Neural Networks With a Full-Precision Auxiliary Module

URL: View paper

Prior Art Analysis

Full-Precision Auxiliary Module[50] demonstrates that auxiliary modules were previously proposed to address gradient propagation difficulties in quantized network training. The candidate paper explicitly introduces learnable auxiliary modules to mitigate quantization errors and stabilize training, predating the original paper's theoretical analysis. Both papers use auxiliary modules to create full-precision routes for gradient updates, though Full-Precision Auxiliary Module[50] focuses on classification/detection while the original targets video generation.

Evidence

Evidence 1 - **Rationale**: Both papers introduce auxiliary modules as a core solution to improve quantized network training, demonstrating prior work on this concept. - **Original**: we introduce auxiliary modules (φ) to mitigate large quantization errors, leading to significantly enhanced convergence. - **Candidate**: we propose a solution by training the low-precision network with a fullprecision auxiliary module. specifically, during training, we construct a mix-precision network by augmenting the original low-precision network with the full precision auxiliary module.

Evidence 2 - **Rationale**: Both papers explicitly state that auxiliary modules create routes to alleviate gradient propagation difficulties caused by quantization. - **Original**: To this end, we introduce auxiliary modules (φ) to mitigate large quantization errors, leading to significantly enhanced convergence. - **Candidate**: the motivation of such a design is to create full-precision routes to update parameters of the low-precision model and thus alleviating the difficulty of propagating gradient in a quantized model.

Evidence 3 - **Rationale**: While the original paper provides theoretical analysis of gradient norm reduction, Full-Precision Auxiliary Module[50] already implemented auxiliary modules that provide more accurate gradient updates through averaging, addressing the same underlying convergence issue. - **Original**: we begin with a theoretical analysis demonstrating that reducing the gradient norm is essential to facilitate convergence for qat. - **Candidate**: the gradient of the shared weights is averaged from both the auxiliary module and the original low-precision network to achieve more accurate updating direction.

Evidence 4 - **Rationale**: Both papers describe how auxiliary modules stabilize optimization by creating additional pathways for gradient flow, though the original adds specific gradient norm analysis. - **Original**: these modules effectively help narrow the discrepancy between the discrete quantized and full-precision models, leading to stable optimization and largely reduced ‖gt‖2. - **Candidate**: this strategy creates additional full-precision routes to update the parameters of the low-precision model, thus making the gradient back-propagates more easily.

3. AQ-DETR: Low-Bit Quantized Detection Transformer with Auxiliary Queries

URL: View paper

Brief Assessment

AQ-DETR[53] introduces auxiliary queries to enhance query capacity in object detection transformers, not auxiliary modules to reduce gradient norm for low-bit quantization convergence in video diffusion models. The technical contexts are fundamentally different.

4. Optimal Clipping and Magnitude-aware Differentiation for Improved Quantization-aware Training

URL: View paper

Brief Assessment

Optimal Clipping QAT[49] focuses on optimizing clipping scalars for quantization noise reduction in QAT, not on introducing auxiliary modules to reduce gradient norms for convergence improvement in video diffusion models.

5. Punching Above Precision: Small Quantized Model Distillation with Learnable Regularizer

URL: View paper

Brief Assessment

Game of Regularizer (GOR)[51] focuses on balancing task-specific and distillation losses through learnable loss weighting, not on introducing auxiliary modules to reduce gradient norm for QAT convergence.

Contribution 3: Rank-decay strategy to eliminate inference overhead

Description: The authors develop a rank-decay strategy that progressively removes auxiliary modules during training to eliminate inference overhead. This strategy repeatedly applies singular value decomposition and rank-based regularization to identify and decay low-contributing components, ultimately achieving zero additional inference cost while maintaining performance.

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. Unveiling lora intrinsic ranks via salience analysis

URL: View paper

Brief Assessment

LoRA Intrinsic Ranks[46] focuses on adaptive rank allocation during LoRA fine-tuning through salience analysis and time-series windows, not on eliminating auxiliary module inference overhead in quantized video diffusion models through SVD-based rank decay.

2. An efficient SVD-based method for image denoising

URL: View paper

Brief Assessment

SVD Image Denoising[43] applies SVD for low-rank approximation in image denoising by selecting largest singular values, not for progressively eliminating auxiliary modules during training to remove inference overhead in quantized video generation models.

3. Structure-Preserving Network Compression Via Low-Rank Induced Training Through Linear Layers Composition

URL: View paper

Brief Assessment

Structure-Preserving Compression[45] focuses on compression-promoted training through linear layer composition for network compression, not on eliminating auxiliary modules during video diffusion model training. The technical contexts and application domains differ fundamentally.

4. Using SVD for topic modeling

URL: View paper

Brief Assessment

SVD Topic Modeling[40] applies SVD for topic modeling in text analysis, not for eliminating auxiliary modules in quantized video generation models. The technical contexts are fundamentally different.

5. Self-supervised knowledge distillation using singular value decomposition

URL: View paper

Brief Assessment

SVD Knowledge Distillation[44] uses SVD for knowledge compression in teacher-student networks but does not employ a rank-decay strategy to progressively eliminate auxiliary modules during training. The candidate focuses on distilling knowledge between teacher and student networks, not on removing inference overhead from auxiliary components.

6. Full-Rank No More: Low-Rank Weight Training for Modern Speech Recognition Models

URL: View paper

Brief Assessment

Low-Rank Speech Recognition[48] focuses on low-rank weight training for speech recognition models from scratch, not on progressively eliminating auxiliary modules during training to remove inference overhead in quantized video diffusion models.

7. Flashsvd: Memory-efficient inference with streaming for low-rank models

URL: View paper

Brief Assessment

Flashsvd[42] focuses on streaming inference for SVD-compressed models to reduce activation memory during inference, not on progressively eliminating auxiliary modules during training to achieve zero inference overhead as in the original paper's rank-decay strategy.

8. Adaptive Rank Allocation for Federated Parameter-Efficient Fine-Tuning of Language Models

URL: View paper

Brief Assessment

Adaptive Rank Allocation[47] focuses on federated learning settings with rank-based module pruning to reduce training costs across distributed clients, not on eliminating auxiliary module inference overhead through progressive SVD-based decay during centralized quantization-aware training.

9. Randomized greedy magic point selection schemes for nonlinear model reduction

URL: View paper

Brief Assessment

Greedy Magic Point[41] focuses on reducing computational costs in model reduction via rank-one SVD updates for magic point selection, not on eliminating auxiliary module inference overhead in neural network training.

10. Accelerated SVD-based initialization for nonnegative matrix factorization

URL: View paper

Brief Assessment

Accelerated SVD Initialization[39] focuses on SVD-based initialization for NMF dimensionality reduction, not on eliminating auxiliary module inference overhead during neural network training. The technical contexts are fundamentally different.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] QVGen: Pushing the Limit of Quantized Video Generative Models View paper
- [1] Tr-dq: Time-rotation diffusion quantization View paper
- [2] TCAQ-DM: timestep-channel adaptive quantization for diffusion models View paper
- [3] PARO: Hardware-Software Co-design with Pattern-aware Reorder-based Attention Quantization in Video Generation Models View paper
- [4] Q-VDiT: Towards Accurate Quantization and Distillation of Video-Generation Diffusion Transformers View paper
- [5] Quantization as a Foundation for Deployable High Performance Diffusion Models within the Landscape of Large Scale Generative AI View paper
- [6] FPSAttention: Training-Aware FP8 and Sparsity Co-Design for Fast Video Diffusion View paper
- [7] ViDiT-Q: Efficient and Accurate Quantization of Diffusion Transformers for Image and Video Generation View paper
- [8] Securing Federated Diffusion Model With Dynamic Quantization for Generative AI Services in Multiple-Access Artificial Intelligence of Things View paper
- [9] QuantCache: Adaptive Importance-Guided Quantization with Hierarchical Latent and Layer Caching for Video Generation View paper
- [10] TaQ-DiT: Time-aware Quantization for Diffusion Transformers View paper
- [11] TFMQ-DM: Temporal Feature Maintenance Quantization for Diffusion Models View paper
- [12] Yume: An Interactive World Generation Model View paper
- [13] LLIA - Enabling Low-Latency Interactive Avatars: Real-Time Audio-Driven Portrait Video Generation with Diffusion Models View paper
- [14] Hardware-Friendly Static Quantization Method for Video Diffusion Transformers View paper
- [15] Bitrate-Controlled Diffusion for Disentangling Motion and Content in Video View paper
- [16] ARLON: Boosting Diffusion Transformers with Autoregressive Models for Long Video Generation View paper
- [17] Q-DiT: Accurate Post-Training Quantization for Diffusion Transformers View paper
- [18] QuantSparse: Comprehensively Compressing Video Diffusion Transformer with Model Quantization and Attention Sparsification View paper
- [19] QVD: Post-training Quantization for Video Diffusion Models View paper
- [20] S^{2}Q-VDiT: Accurate Quantized Video Diffusion Transformer with Salient Data and Sparse Token Distillation View paper
- [21] DVD-Quant: Data-free Video Diffusion Transformers Quantization View paper
- [22] LRQ-DiT: Log-Rotation Post-Training Quantization of Diffusion Transformers for Image and Video Generation View paper
- [23] FraQAT: Quantization Aware Training with Fractional bits View paper
- [24] Adaptive Compression and Quantization Techniques for Robust and Scalable Generative Diffusion Networks View paper
- [25] Advanced Sign Language Video Generation with Compressed and Quantized Multi-Condition Tokenization View paper
- [26] Low-Bit Generative Modeling with Diffusion Networks for Scalable and Perception-Aware Synthesis View paper
- [27] Quantizing Diffusion Models for Scalable and Efficient Generative Inference Across Diverse Hardware Platforms View paper
- [28] Temporal Dynamic Quantization for Diffusion Models View paper
- [29] From High Precision Denoising to Lightweight Generation with Quantized Diffusion Models View paper
- [30] CLQ: Cross-Layer Guided Orthogonal-based Quantization for Diffusion Transformers View paper
- [31] DilateQuant: Accurate and Efficient Quantization-Aware Training for Diffusion Models via Weight Dilation View paper
- [32] Contemporary Advances in Neural Network Quantization: A Survey View paper
- [33] Quantizing Diffusion Models from a Sampling-Aware Perspective View paper
- [34] Dynamic and Mixed-Precision Techniques for Scalable Iterative Generative Modeling View paper
- [35] UniHM: Universal Human Motion Generation with Object Interactions in Indoor Scenes View paper
- [36] DilateQuant: Accurate and Efficient Diffusion Quantization via Weight Dilation View paper
- [37] GFix: Perceptually Enhanced Gaussian Splatting Video Compression View paper
- [38] VETA-DiT: Variance-Equalized and Temporally Adaptive Quantization for Efficient 4-bit Diffusion Transformers View paper
- [39] Accelerated SVD-based initialization for nonnegative matrix factorization View paper
- [40] Using SVD for topic modeling View paper
- [41] Randomized greedy magic point selection schemes for nonlinear model reduction View paper
- [42] Flashsvd: Memory-efficient inference with streaming for low-rank models View paper
- [43] An efficient SVD-based method for image denoising View paper
- [44] Self-supervised knowledge distillation using singular value decomposition View paper
- [45] Structure-Preserving Network Compression Via Low-Rank Induced Training Through Linear Layers Composition View paper
- [46] Unveiling lora intrinsic ranks via salience analysis View paper
- [47] Adaptive Rank Allocation for Federated Parameter-Efficient Fine-Tuning of Language Models View paper
- [48] Full-Rank No More: Low-Rank Weight Training for Modern Speech Recognition Models View paper
- [49] Optimal Clipping and Magnitude-aware Differentiation for Improved Quantization-aware Training View paper

- [50] Training Quantized Neural Networks With a Full-Precision Auxiliary Module View paper
- [51] Punching Above Precision: Small Quantized Model Distillation with Learnable Regularizer View paper
- [52] Stable Quantization-Aware Training with Adaptive Gradient Clipping View paper
- [53] AQ-DETR: Low-Bit Quantized Detection Transformer with Auxiliary Queries View paper
- [54] Bidm: Pushing the limit of quantization for diffusion models View paper
- [55] PTQ4DiT: Post-training Quantization for Diffusion Transformers View paper
- [56] Q-dm: An efficient low-bit quantized diffusion model View paper
- [57] MPQ-DM: Mixed Precision Quantization for Extremely Low Bit Diffusion Models View paper
- [58] MPQ-DMv2: Flexible Residual Mixed Precision Quantization for Low-Bit Diffusion Models with Temporal Distillation View paper