Dynamic and Quantization-Aware Attention Reuse for Diffusion Transformers

Project Proposal

Name: Gautham Satyanarayana **Date:** October 17, 2025

1. Introduction

Diffusion Transformers (DiTs) have become a dominant architecture for large-scale generative modeling, combining the scalability of transformers with the denoising diffusion process. However, their inference remains costly due to quadratic self-attention and large key–value (KV) caches that are recomputed for each timestep.

Recent work, including *Attention Compression for Diffusion Transformer Models* [1], demonstrated that attention maps exhibit strong temporal redundancy during sampling, allowing for partial reuse between steps. In parallel, *PTQ4DiT* [2] and *Q-DiT* [3] addressed quantization challenges unique to diffusion transformers—specifically, salient activation channels and timestep-dependent distributions. Building on these advances, this project proposes a unified framework, **Dynamic and Quantization-Aware Attention Reuse (DQAR)**, that combines entropy- and SNR-based attention reuse with low-bit quantized KV caching.

2. Related Work

2.1. Attention Compression and Reuse

[1] proposed three strategies—Window Attention with Residual Sharing, Attention Sharing Across Timesteps, and CFG Branch Sharing—to exploit redundancy in DiTs. Their method achieved significant FLOP reductions while maintaining image quality, showing that attention maps can be reused safely once convergence begins.

2.2. Quantization in Diffusion Transformers

PTQ4DiT [2] introduced channel-wise salience balancing (CSB) and Spearman's ρ–guided calibration to stabilize post-training quantization across timesteps. *Q-DiT* [3] extended this by proposing automatic granularity allocation and dynamic activation quantization. Outside DiTs, *EfficientDM* [4] explored quantization-aware fine-tuning for general diffusion models, showing that hybrid quantization can yield strong efficiency–quality trade-offs.

2.3. Motivation for DQAR

Existing approaches treat compression and quantization independently. Our framework integrates both: (1) using information-theoretic signals (entropy and SNR) to decide when to reuse cached attention; and (2) quantizing reused KV tensors to minimize VRAM and I/O cost. This coupling creates an adaptive, compute-efficient inference pipeline for DiTs.

3. Design and Methodology

3.1. 1. Entropy & SNR-Based Reuse Gate

At each timestep t, we compute the attention entropy

$$H_t = -\frac{1}{HT} \sum_{h=1}^{H} \sum_{i,j} A_t^{(h)}(i,j) \log(A_t^{(h)}(i,j) + \epsilon)$$

and latent signal-to-noise ratio

$$\text{SNR}_t = \frac{\|x_0\|_2^2}{\|x_t - x_0\|_2^2 + \epsilon}.$$

Attention is reused only if $H_t < \tau(p)$ and $\text{SNR}_t \in [a, b]$, where $\tau(p)$ adapts to prompt length.

3.2. 2. Quantized KV Caching

Cached K/V tensors are stored in 8-bit integer form:

$$K_q = \text{clip}\left(\text{round}\left(\frac{K}{s_K}\right), -127, 127\right), \quad s_K = \frac{\max|K|}{127}.$$

Two modes are supported: (1) per-tensor scaling for speed, and (2) per-channel scaling for fidelity. A mixed-precision variant keeps K in FP16 and quantizes V.

3.3. 3. Dynamic Reuse Policy

A lightweight MLP (< 0.5M parameters) predicts the probability of reuse:

$$p_{\text{reuse}} = P_\theta(\text{concat}(H_t, \text{SNR}_t, \|x_t\|_2, t)).$$

The policy is trained offline on cached inference traces labeled by performance impact, enabling data-driven decisions without modifying the diffusion model.

3.4. 4. Layer Scheduling

Early timesteps reuse only shallow blocks (where entropy is high and signal weak), while later timesteps reuse deeper blocks (where convergence stabilizes attention). This scheduling aligns reuse intensity with the diffusion timeline.

3.5. 5. Integration

DQAR modules integrate directly into standard DDIM or DPMSolver samplers. Each diffusion step:

1. Computes entropy and SNR for the current step.
2. Consults the reuse gate or policy.
3. If approved, retrieves quantized K/V and dequantizes them.
4. Otherwise, recomputes attention and updates caches.

4. Experimental Setup

- **Base Model:** DiT-Base or DiT-XL (Hugging Face / OpenDiT)
- **Hardware:** A100 GPU (Google Colab Pro)
- **Data:** 64–256 text prompts from COCO-2017 or LAION
- **Baselines:** FP16 DiT, Attention Compression (static), PTQ4DiT (quantized)
- **Metrics:** FID, CLIP score, runtime, VRAM usage
- **Ablations:** Reuse off / Entropy gate / Quant cache / Policy head

5. Expected Outcomes

- Achieve $\geq 25\%$ inference speedup or $\geq 20\%$ VRAM reduction with ≤ 1 FID degradation.
- Empirically map the relationship between attention entropy, SNR, and quality stability.
- Release the first open-source code unifying reuse and quantization for DiTs.

References

[1] Yuan, Zhihang; Zhang, Hanling; Pu, Lu; Ning, Xuefei; Zhang, Linfeng; Zhao, Tianchen; Yan, Shengen; Dai, Guohao; Wang, Yu. **Attention Compression for Diffusion Transformer Models**. In *NeurIPS*, 2024.

[2] Wu, Junyi; Wang, Haoxuan; Shang, Yuzhang; Shah, Mubarak; Yan, Yan. **PTQ4DiT: Post-training Quantization for Diffusion Transformers**. In *NeurIPS*, 2024. https://arxiv.org/abs/2405.16005

[3] Chen, Lei; Meng, Yuan; Tang, Chen; Ma, Xinzhu; Jiang, Jingyan; Wang, Xin; Wang, Zhi; Zhu, Wenwu. **Q-DiT: Accurate Post-Training Quantization for Diffusion Transformers**. *arXiv preprint arXiv:2406.17343*, 2024.

[4] He, Yefei; Liu, Jing; Wu, Weijia; Zhou, Hong; Zhuang, Bohan. **EfficientDM: Efficient Quantization-Aware Fine-Tuning of Low-Bit Diffusion Models**. *arXiv preprint arXiv:2310.03270*, 2023.