

Peer Review

Review of: "Efficiency Meets Fidelity: A Novel Quantization Framework for Stable Diffusion"

Yuangang Pan¹¹. Agency for Science, Technology and Research, Singapore, Singapore

The study of model quantization for diffusion models, especially the inconsistency issue in both the calibration and inference processes in existing works, is a practical problem. The proposed mixed-precision quantization strategy, multi-timestep activation quantization, and time information precalculation techniques sound reasonable and interesting, with experimental verification. However, the presentation of the technique is poor and requires further improvement:

1. The positioning of the paper is unclear. The comparison with related works, especially the baselines Qdillusion, PTQ4DM, and PCR, is lacking, which makes the evaluation of the contribution difficult.
2. It seems Eq. 12 does not depend on S_i ; more details should be provided here. The detailed formation of Eq. 12 should be provided.
3. The presentation of this paper is poor, and some words are confusing. For example, which terms in Eq. 9 are redundant, and why are they called redundant? How is the sensitivity ranking performed?
4. There are two lines marked as "ours" in Table 4 and Table 5; what is the difference?
5. The caption of Figure 6 may be wrong.

Declarations

Potential competing interests: No potential competing interests to declare.