LAQFI

Layer-Adaptive Quantization on Diffusion Model using Fisher Information

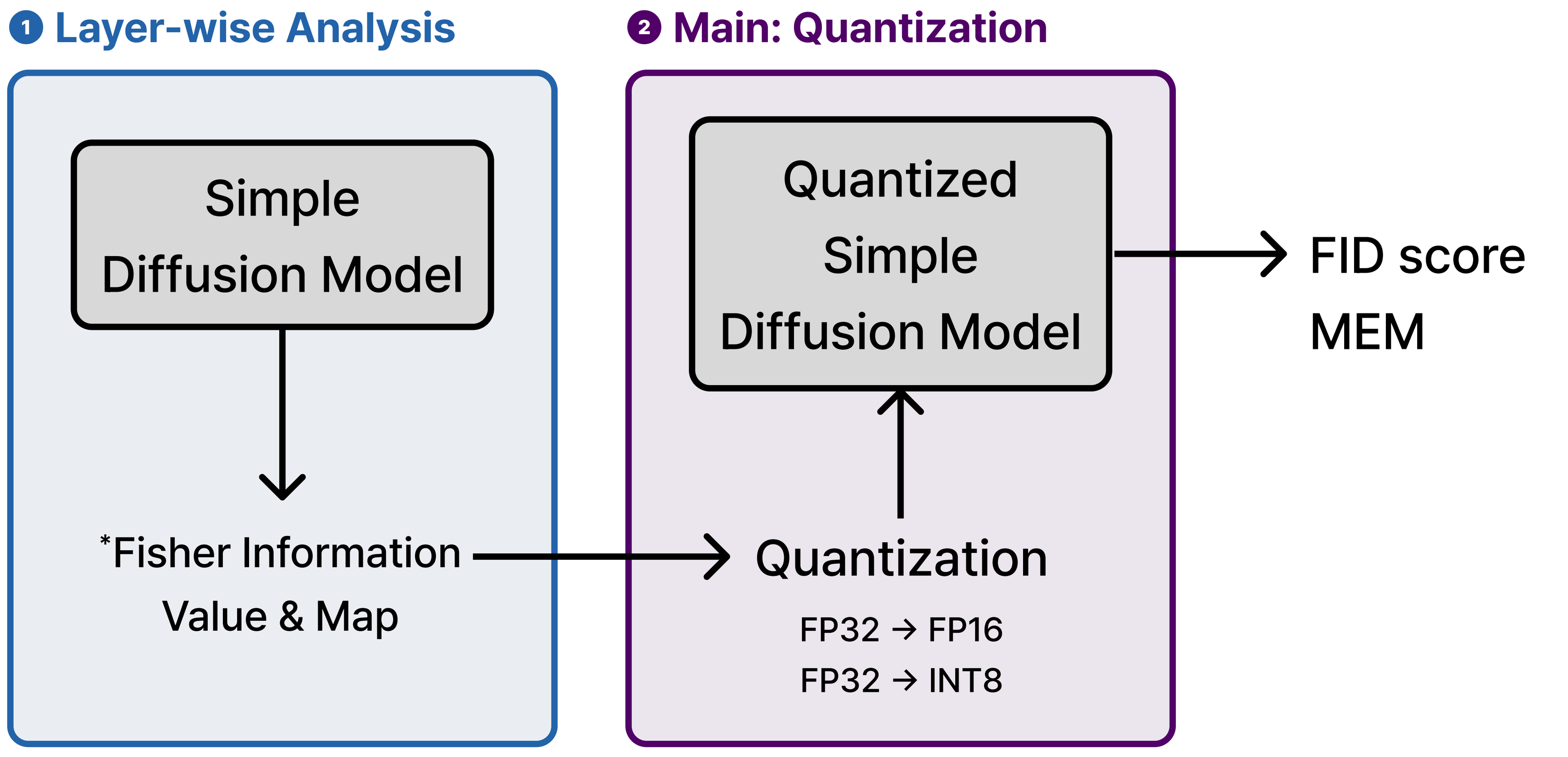
Doeun Kim[1], Jieun Byeon[1], Inae Park[1], Jaehyeong Sim[*]

Background and Overview

Diffusion models deliver outstanding generative performance with high-quality images, but their large parameter sizes lead to substantial memory usage. Quantization has emerged as a key approach to improve storage and deployment efficiency. However, most existing works apply a uniform bit-width across the entire model, often resulting in noticeable quality degradation.

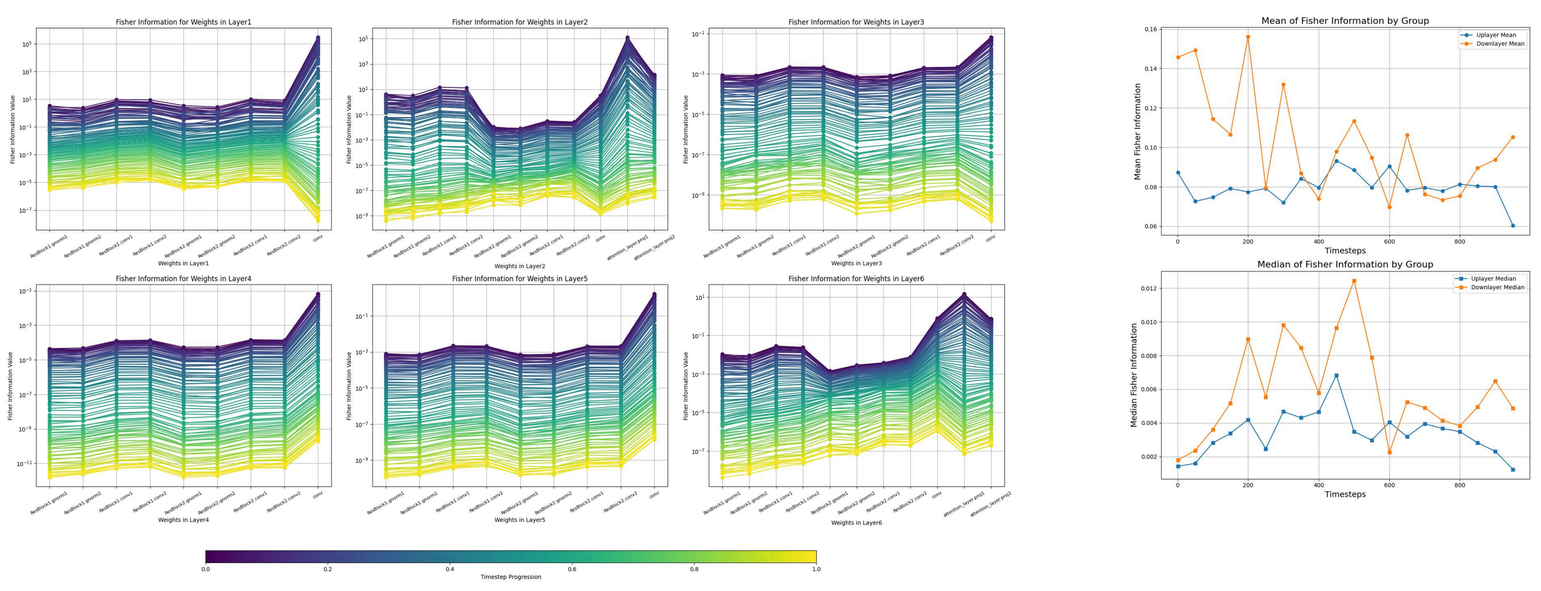
In this study, we propose **Layer-Adaptive Quantization using Fisher Information**, which **leverages Fisher Information to quantify each layer's contribution and importance**, and applies **differentiated precision levels based on per-layer significance** rather than uniform quantization.

Pipeline



* The Fisher Information Matrix quantifies parameter sensitivity and importance.
We use it to assess each layer's contribution across time steps in diffusion models.

Step 1: Layer-wise Analysis



▲ Fisher Information per layer across timesteps* (Left) and by layer groups (Layer 1–2 / 3–6) (Right)

- Later timesteps show generally higher weight contributions.
- Fisher Information distributions vary by layer, with Layer 1 and 2 showing notably high contribution to image generation.

* Log-scale normalization applied.
* Values closer to 1 (yellow) indicate earlier timesteps; closer to 0 (purple) indicate later timesteps.

Step 2: Quantization Through Threshold Setting

[Baseline] Full-precision model

Methods

We train a Simple Diffusion Model on MNIST and generate 100 images over 1,000 timesteps, evaluating FID, IS, and memory footprint.

Results

- FID : **59.3030**
- IS: **1.7225 ± 0.1775**
- MEM : 134.20 MB

1) Global Thresholding

Methods

We apply a single threshold to Fisher Information values across all layers during the backward process of the Simple Diffusion Model.

Results

- FID: **59.7148**
- IS: 2.0846 ± 0.3764
- MEM: **71.48 MB**

2) Group-wise Thresholding

Methods

Based on the layer-wise Fisher Analysis, we divide the model into three layer groups (Layer 1–2 / Layer 3–5 / Layer 6) and assign a different threshold to each group.

Results

- FID: **57.8034**
- IS: 1.7356 ± 0.1941
- MEM: 134.

3) Layer-wise Thresholding ❶

Methods

For each of the six layers, a threshold is adaptively determined using statistical measures (mean and standard deviation). We quantize weights whose Fisher values fall below the threshold.

threshold_min = layer_mean - k * layer_std
threshold_max = layer_mean + k * layer_std

Results

- FID Score: **55.5387**
- MEM: **81.56 MB**

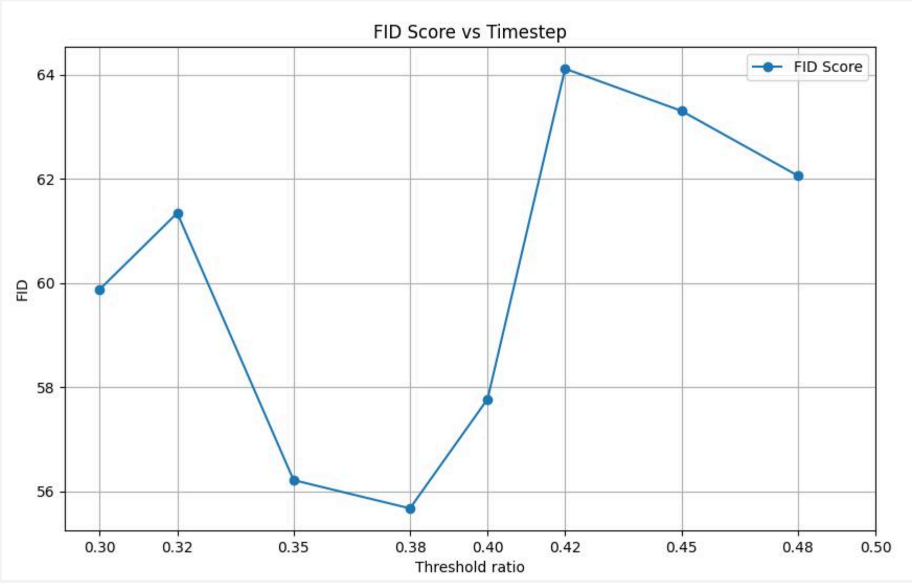
3) Layer별 임계값 설정❷ - 임계값 비율

Methods

We set a threshold ratio* for each of the six layers in the Simple Diffusion Model to compute individual thresholds and quantize weights whose Fisher values fall below those thresholds.

* Threshold ratio: a value between 0.0–1.0 used to define the layer-wise threshold (e.g., 0.2 → top 20% Fisher values).

Results



Best performance appears in the 0.35–0.40 threshold ratio range, with 0.35 yielding the optimal balance of FID, IS, and memory usage.

threshold_ratio	FID	IS	Memory
Original (w/o quant)	59.3030	1.7225	134.20 MB
0.35	56.2130	1.7628	68.02 MB
0.38	55.6704	1.8436	64.23 MB
0.40	57.7688	1.7454	67.70 MB

Conclusion and Contributions

Unlike conventional quantization methods, our approach incorporates **layer-wise importance analysis** into the quantization process. Experiments demonstrate that **setting different thresholds for each layer** results in **higher memory compression efficiency and superior image quality** compared to uniform quantization. These results indicate that our method provides a better balance between compactness and generative performance in diffusion models, and can be further extended with more fine-grained threshold assignment strategies.