A. Appendix

Visualization and sampling results form DiTAS In this section, the visualization and sampling outcomes from DiTAS are presented, highlighting the excellent performance of DiTAS under diverse conditions.

Figure 1 and 3 shows the distribution of the activation matrix at a specific time step across each input channel. The outliers across input channels are very explict. After ultilizing our proposed temporal-aggregated smoothing (TAS) and layer-wise grid search optimization strategy, the outliers can be mitigated greatly, as shown in Figure 2 and 4

As shown in 5, 6, 7, and 8 illustrate the sample images rendered by DiTAS under varying conditions, including different image resolutions (512x512 and 256x256 pixels), sampling steps, and precision configurations: W8A8 (8-bit weight and 8-bit activation) and W4A8 (4-bit weight and 8-bit activation). It is clear that DiTAS can produce images with quality comparable to full precision generative images across various bit-width settings.

As shown in Figures 9, 10, 11, 12, and 13, we can find out directly applying the SmoothQuant method to DiT by selecting calibration data from a single time-step could potentially degrade the performance of the quantized DiT. Thus it can demonstrate the effectiveness of our proposed temporal-aggregated smoothing (TAS) strategy.

Code implementation of DiTAS Code is available at https://github.com/DZY122/DiTAS

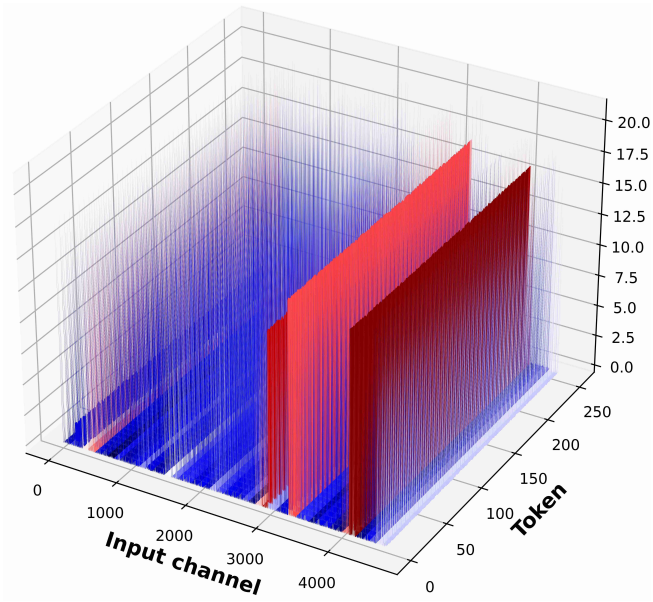

Figure 1. Activation with outliers across input channels in 27th DiT Block's FC2 layer before Temporal-aggregated Smoothing (TAS).

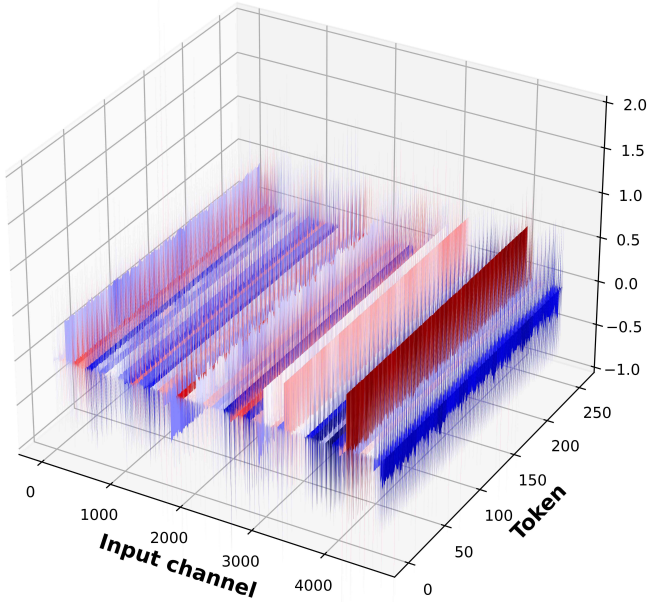

Figure 2. Activation in 27th DiT Block's FC2 layer after Temporal-aggregated Smoothing (TAS) and grid search optimization.

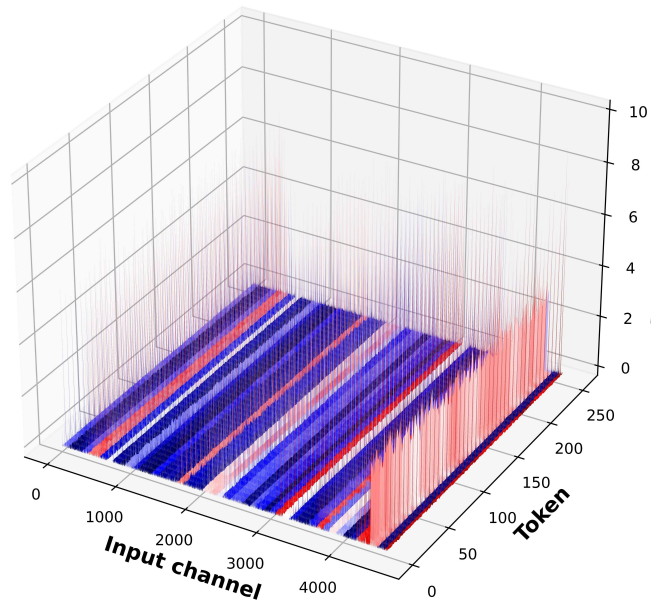

Figure 3. Activation with outliers across input channels in 26th DiT Block's FC2 layer before Temporal-aggregated Smoothing (TAS).

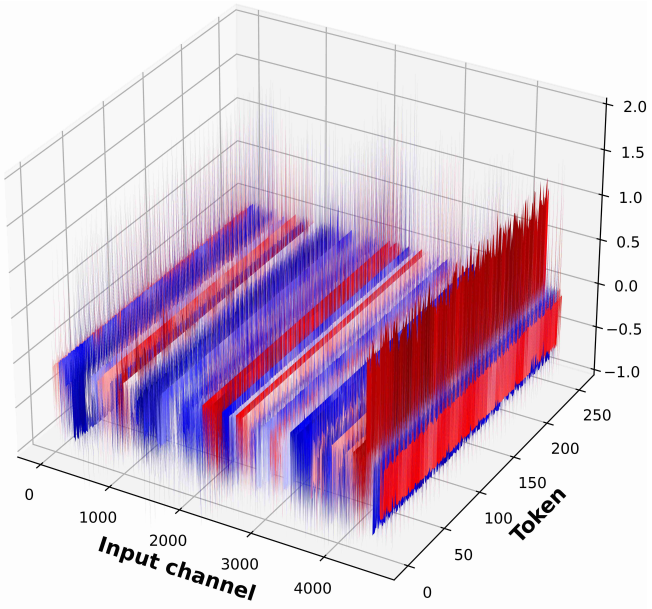

Figure 4. Activation in 26th DiT Block's FC2 layer after Temporal-aggregated Smoothing (TAS) and grid search optimization.

Figure 5. Samples generated by W4A8 DiTAS on ImageNet 256×256 (cfg=4.0, step=50).

Figure 6. Samples generated by W8A8 DiTAS on ImageNet 256×256 (cfg=4.0, step=50).

Figure 7. Samples generated by W4A8 DiTAS on ImageNet 512×512 (cfg=4.0, step=50).

Figure 8. Samples generated by W4A8 DiTAS on ImageNet 512×512 (cfg=4.0, step=100).

Figure 9. Samples generated by naive quantized DiT on ImageNet 256×256 (W4A8, cfg=4.0, step=50).

Figure 10. Select time-step 1 to operate SmoothQuant on 256×256 (W4A8, cfg=4.0, step=50).

Figure 11. Select time-step 25 to operate SmoothQuant 256×256 (W4A8, cfg=4.0, step=50).

Figure 12. Select time-step 50 to operate SmoothQuant 256×256 (W4A8, cfg=4.0, step=50).

Figure 13. Adopt temporal-aggregated smoothing (TAS) on ImageNet 256×256 (W4A8, cfg=4.0, step=50).