Re-ttention: Ultra Sparse Visual Generation via Attention Statistical Reshape

Ruichen Chen[1], Keith G. Mills[2], Liyao Jiang[1], Chao Gao[3], Di Niu[1]

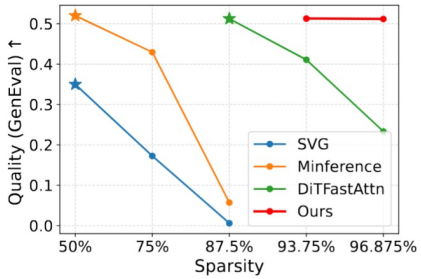
[1]University of Alberta, [2]Louisiana State University, [3]Huawei Technologies

Sparse Attention

- Computes attention only for a selected subset of tokens, skipping less important pairs.
- Reduces the quadratic complexity of full attention to near-linear levels.
- Sparser attention leads to greater computational efficiency.

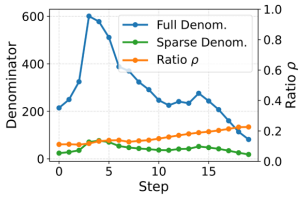
Challenge

- *Preliminary*: Sparsity is defined as the proportion of attention computations skipped by only considering a subset of query-key interactions.
- Existing method achieves at most 87.5% sparsity. However, it is challenging to achieve **ultra sparse** attention, i.e. sparsity greater than 90%.



Key Insight

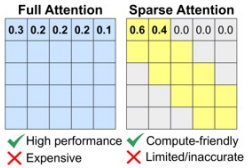
The ratio ρ between the full Softmax denominator and the sparse denominator is relatively stable



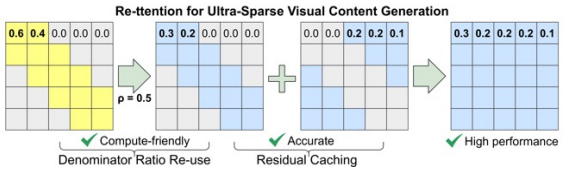
Reshape the sparse attention can approximate full attention more faithfully

Attention Statistical Reshape

Sparse attention normalizes over only a subset of tokens, which alters the Softmax denominator and shifts the distribution of the normalized attention scores away from that of full attention.



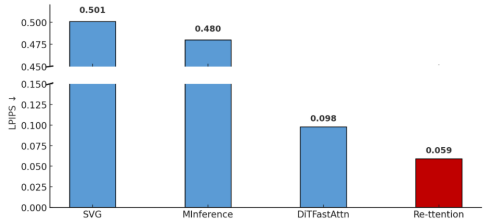
We cache residual R with denominator ratio ρ at the caching timestep. At the subsequent steps, we reshape the sparse attention based on the cached statistics.



Experimental Result

T2V result on CogVideoX-2B.

LPIPS (similarity) metric on CogVideoX-2B.



T2I result on PixArt and Hunyuan DiT.

Figure 6: Visual comparison on MS-COCO 2014 [25] prompts using PixArt-α (row 1), PixArt-Σ (row 2), and Hunyuan (row 3). We show images generated by Re-ttention (our method) and by other attention methods in different columns. We provide further examples in the appendix.