Efficient Long-Context Modeling in Diffusion Language Models via Block Approximate Sparse Attention

Supplementary Material

A. More Ablation Studies

Table 8. Ablation on covariance compensation coefficient β. Results on RealWorldQA and RULER4K (Acc.%).

β	VideoMME	RealWorldQA	RULER4K
0.0	55.82	64.62	88.42
0.1	56.19	64.32	**91.11**
0.5	**56.56**	**64.84**	90.06
1.0	56.51	64.84	86.12
2.0	51.42	62.83	80.1

Effect of the covariance compensation coefficient. Table 8 reports the ablation results on the covariance compensation coefficient β over three representative benchmarks: VideoMME for long-context video understanding, RealWorldQA for real-world image–language reasoning, and RULER4K for ultra-long NLP sequences. When β=0.0, the model degenerates to a pure block-mean baseline without any second-order correction, yielding the lowest accuracy on all three datasets. Introducing a small amount of covariance correction already helps: β=0.1 notably boosts RULER4K accuracy (from 88.42% to 91.11%), suggesting that even weak second-order information is beneficial for very long sequences.

Increasing β to 0.5 further improves multimodal performance, achieving the best results on VideoMME and the highest overall average across the three benchmarks. This setting provides a good trade-off between bias removal and numerical stability, so we adopt β=0.5 as the default in our main experiments. In contrast, larger values ($\beta \geq 1.0$) start to over-amplify variance noise: while RealWorldQA remains flat at β=1.0, performance on RULER4K drops sharply, and β=2.0 leads to clear degradation on all three datasets. Overall, moderate covariance compensation yields robust gains, whereas overly aggressive correction undermines stability and accuracy.

Effect of the norm-based token sorting. To illustrate the effect of our norm-based token sorting, we visualize attention maps before and after applying the sorting procedure in Fig. 5. Without sorting, high-magnitude attention scores are scattered across distant positions, making it difficult for block-sparse attention to discover a compact set of informative blocks and forcing it to behave closer to dense attention. After sorting keys within each segment according to their norm statistics, tokens with similar activation strength are rearranged into more contiguous regions. This induces a clear structure in the attention map, where salient interactions are concentrated in a few blocks while many other blocks become largely homogeneous and low-impact. As a result, block-sparse schemes can drop a substantial fraction of blocks with negligible loss of information, allowing our method to preserve most task-relevant signal while significantly reducing computation, especially in early layers where token representations are more diverse.

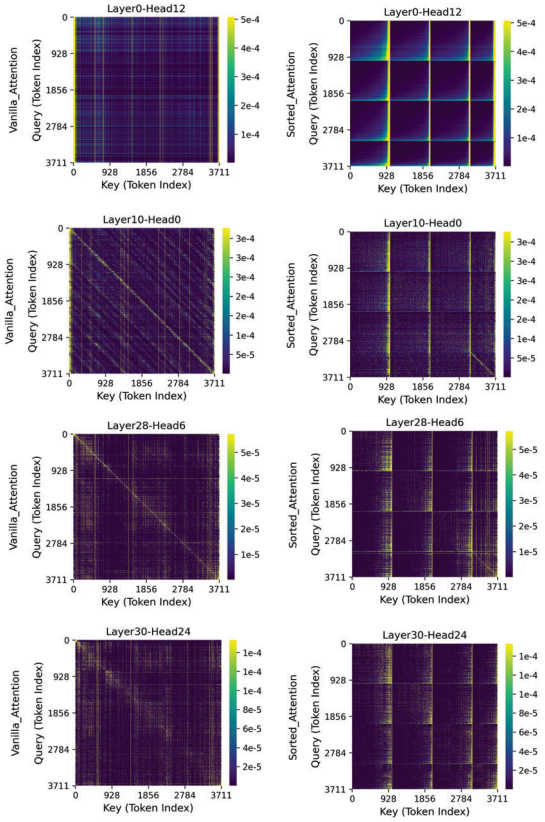

Figure 5. Attention maps before and after norm-based segment sorting. Rows correspond to query tokens, columns to key tokens. After sorting, high-norm (informative) keys are grouped toward the left, enabling sparse attention to better capture salient regions.