NEURAL INFORMATION PROCESSING SYSTEMS

DuSA: Fast and Accurate Dual-Stage Sparse Attention Mechanism Accelerating Both Training and Inference

Chong Wu*[1], Jiawang Cao*[2], Renjie Xu*[1], Zhuoheng Ran[1], Maolin Che†[3], Wenbo Zhu[2], and Hong Yan[1]

*Equal contribution
†Corresponding author (chncml@outlook.com)

1 2 3

➢ Overview of DuSA

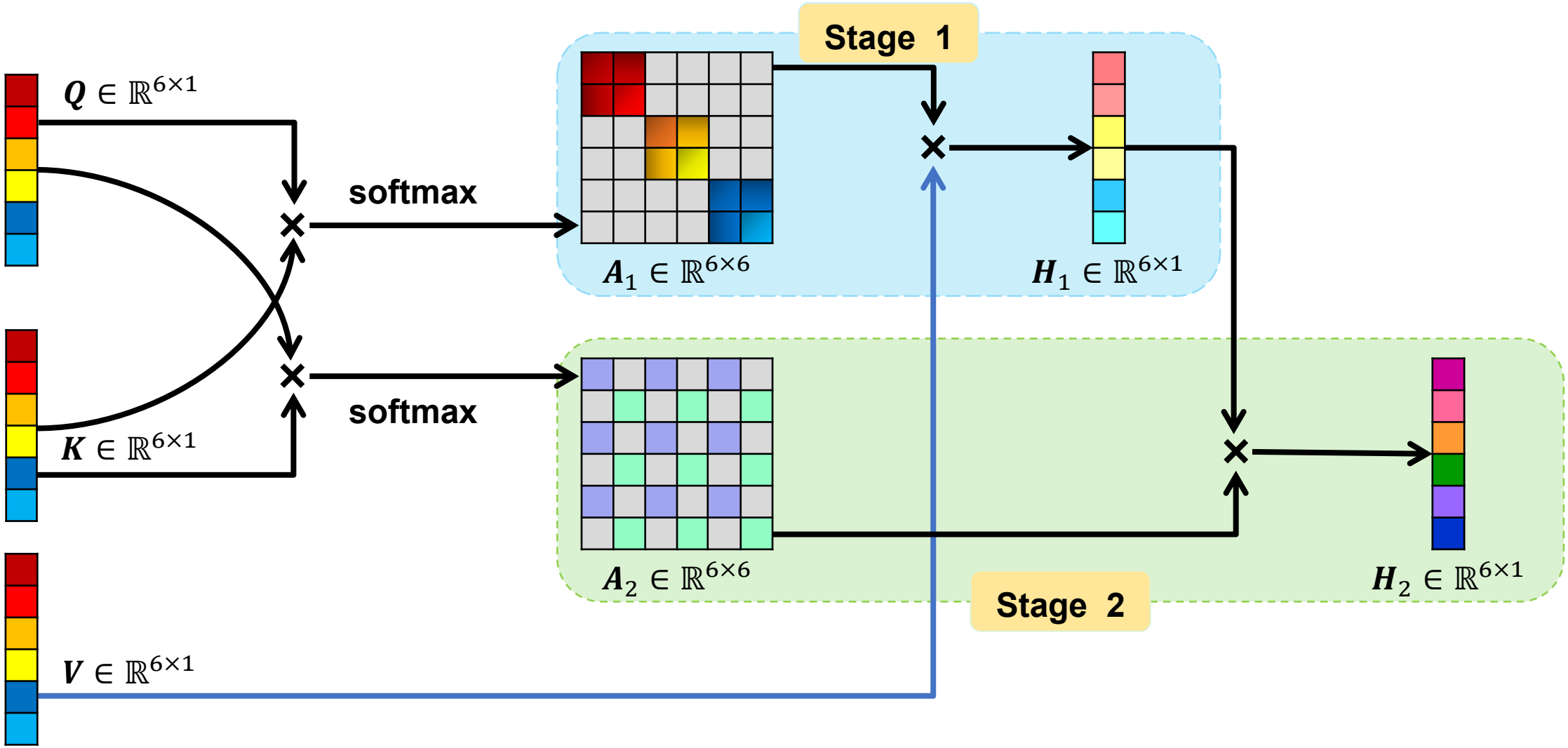
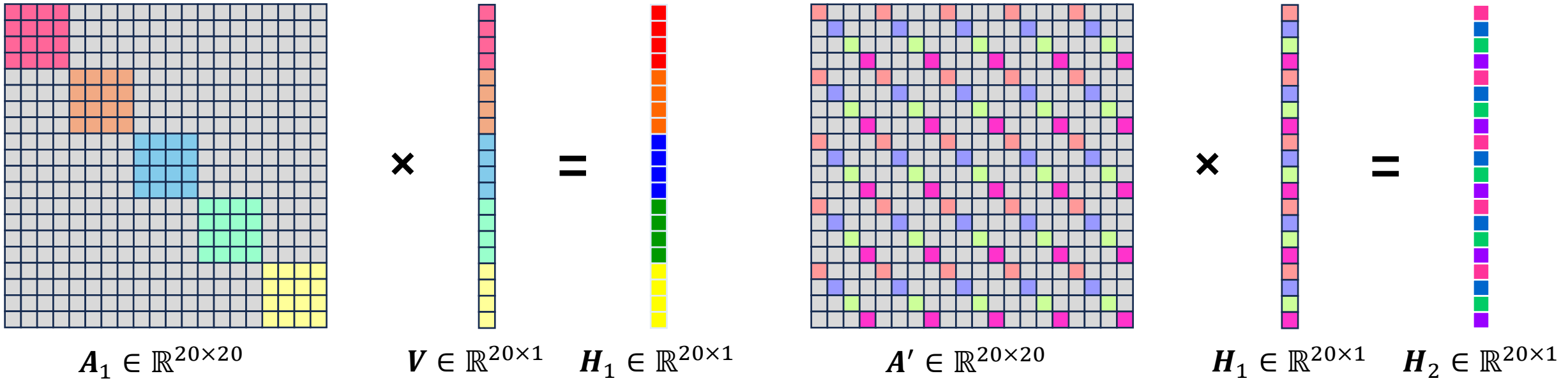


Fig. 1: Overview of DuSA.

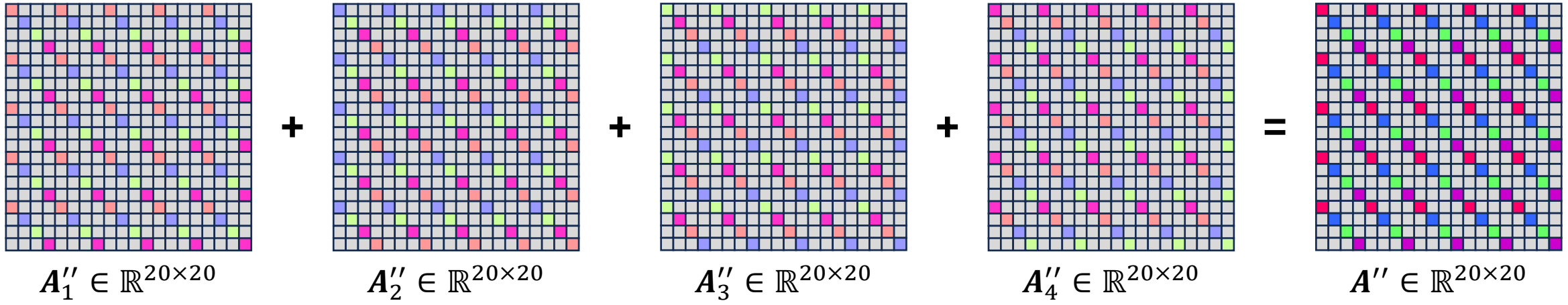
➢ The Dual-Stage Sparse Attention Design



Stage 1: Intrablock Sparse Attention **Stage 2: Interblock Sparse Attention (Strategy 1)**

Fig. 2: In Stage 1, intrablock sparse attention performs token mixing within each diagonal block. In Stage 2, inter-block sparse attention uses **Strategy 1**: **Diagonal Pattern-Based Block Mixing** ($A_2 = A'$) to mix different diagonal block information for each token.

➢ Rolling Indices-Based Attention Scores Fusion for Block Mixing



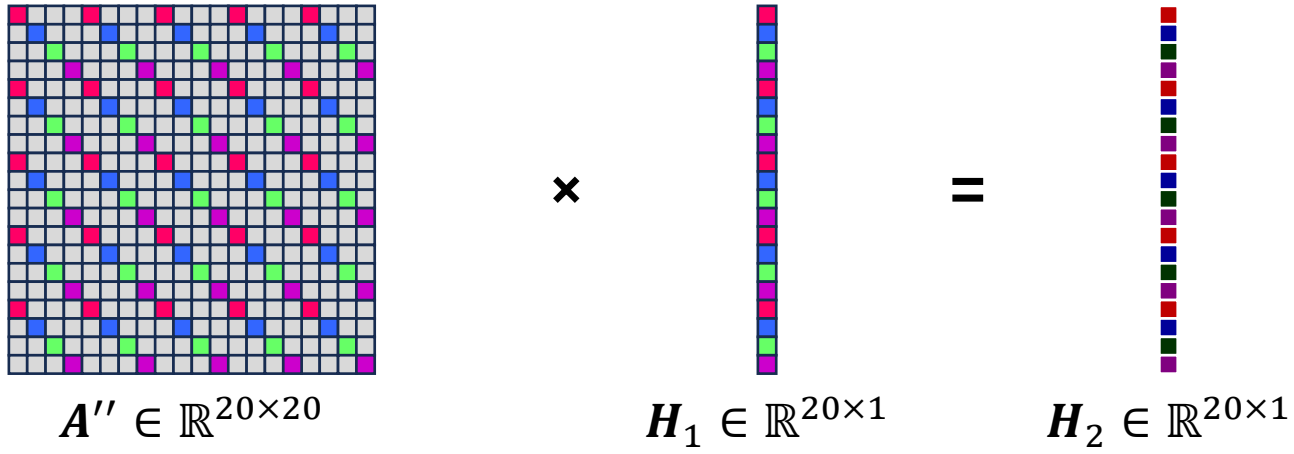
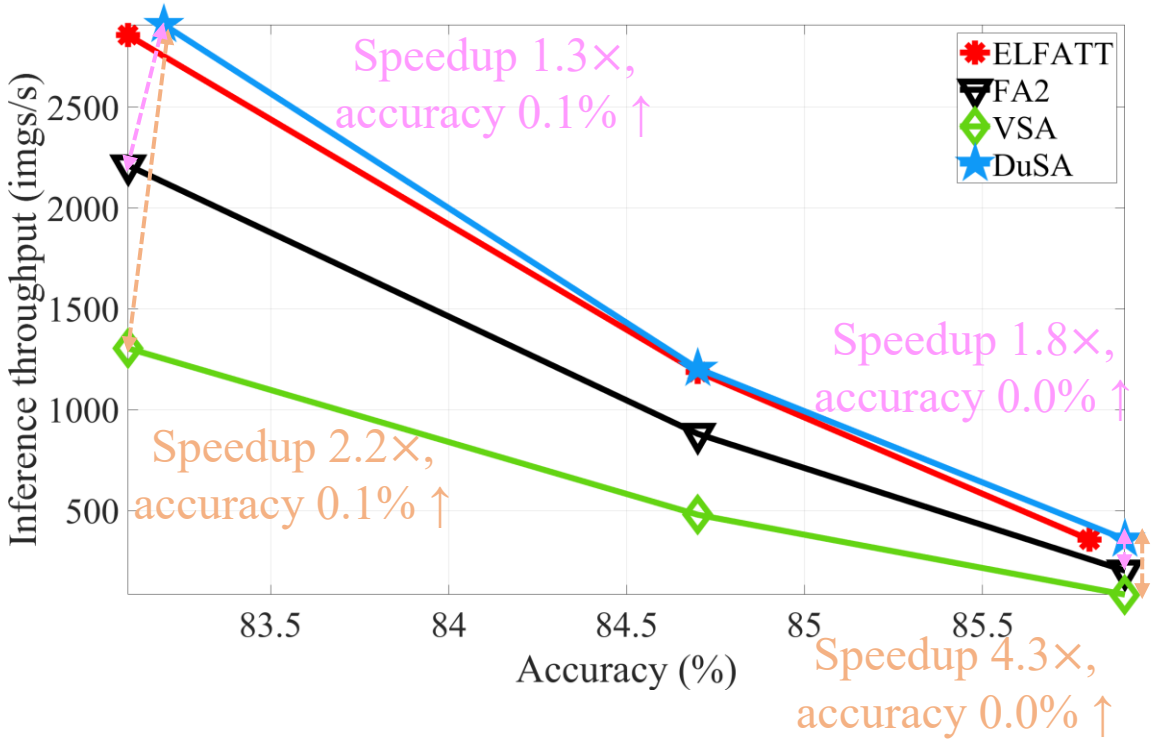
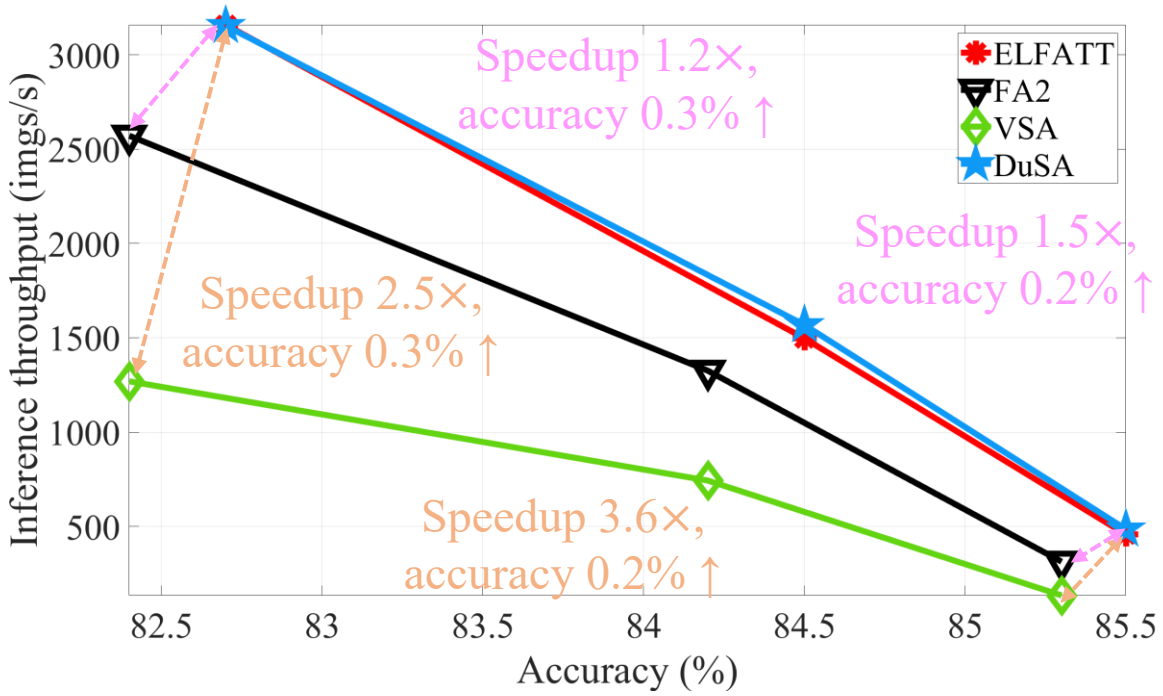


Fig. 3: Stage 2 using **Strategy 2**: **Rolling Indices-Based Attention Scores Fusion for Block Mixing** ($A_2 = A''$).

➢ Accuracy-Efficiency Comparison with SOTA Methods



(a) CSWin-T/B



(b) Swin-T/B

Fig. 4: Accuracy-efficiency curves obtained by different attention mechanisms (ELFATT [1], DuSA, FA2 [2], and VSA [3]) on ImageNet-1K.

➢ Visual Comparison of Class Activation Map (CAM) Based Attention Results

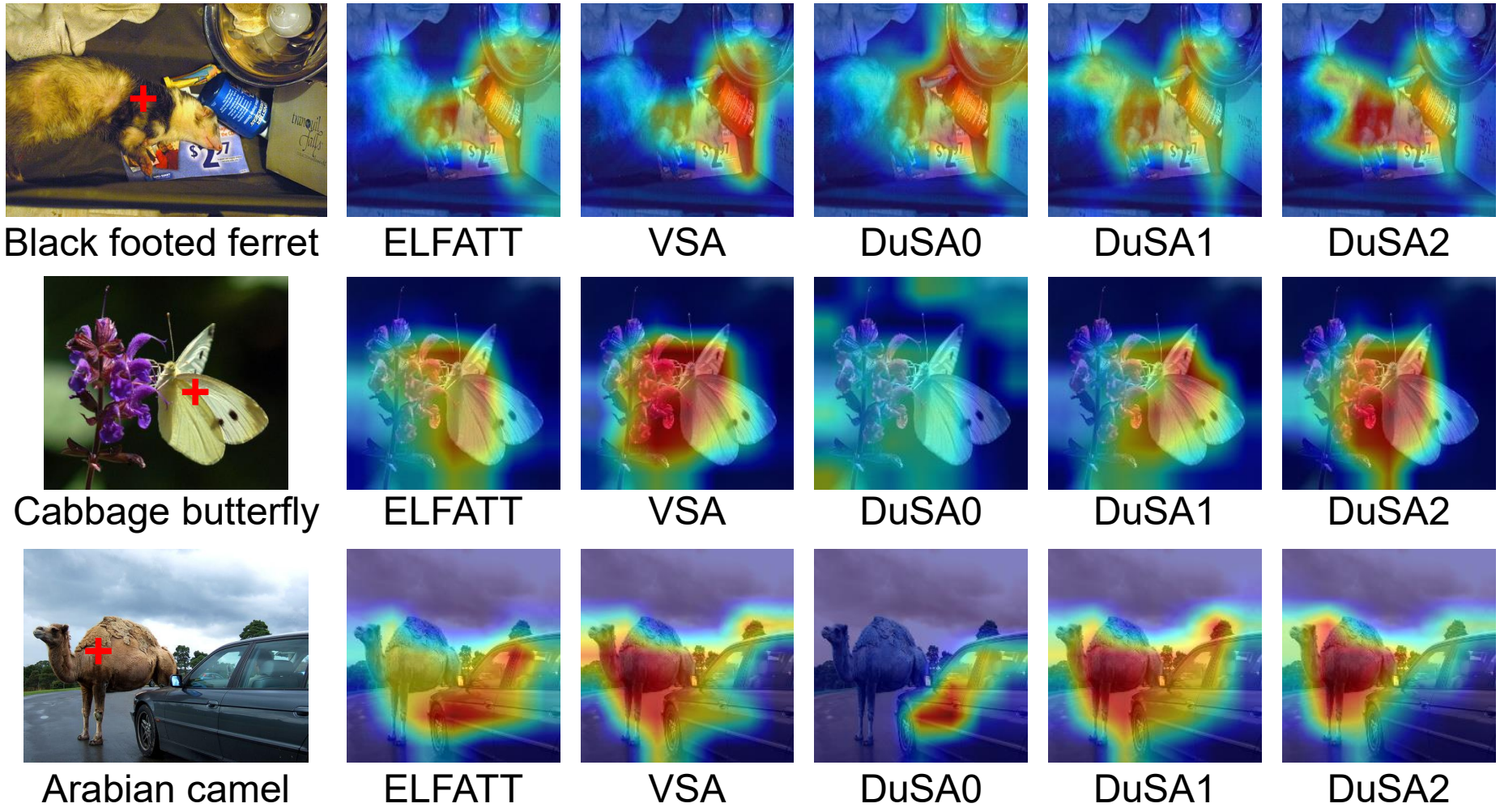


Fig. 5: Visual comparison of class activation map (CAM) based attention results of different attention mechanisms. DuSA0/DuSA1/DuSA2 is DuSA (without using Strategies 1 and 2)/(Strategy 1)/(Strategy 2).

➢ Other Downstream Tasks Comparison

NEURAL INFORMATION PROCESSING SYSTEMS

Tab. 1: Performance comparison on MS COCO 2017.

Method	AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}	FPS (imgs/s)	#
Mask R-CNN 3×MS schedule, Backbone: CSWin-T								
ELFATT	**49.4**	**70.9**	**54.4**	44.0	**68.0**	47.5	33	40M
VSA	48.8	70.0	53.5	43.6	67.4	47.1	7/18 (FA2)	40M
DuSA	**49.4**	**70.9**	**54.4**	**44.1**	67.8	**47.6**	**39**	40M
Mask R-CNN 3×MS schedule, Backbone: Swin-T								
ELFATT	**48.5**	**70.4**	**53.4**	43.6	67.3	47.3	45	50M
VSA	48.0	70.0	52.7	43.3	67.0	46.8	6/17 (FA2)	48M
DuSA	48.4	70.3	53.0	**43.8**	**67.5**	**47.5**	**50**	50M

Tab. 2: Performance comparison on ADE20K.

Method	mAcc	mIoU	FPS (imgs/s)	#
UperNet 160K, Backbone: CSWin-T				
ELFATT	61.2	**49.6**	**32**	50M
VSA	61.1	48.8	6/14 (FA2)	50M
DuSA	**61.4**	**49.6**	**32**	50M
UperNet 160K, Backbone: Swin-T				
ELFATT	59.3	47.7	**38**	62M
VSA	59.3	47.8	5/14 (FA2)	60M
DuSA	**59.8**	**48.0**	37	62M

➢ Text to Video Generation Using CogVideoX-2B on Open-Sora 2.0 Prompt Sets

FA2 (CLIPSIM: 0.1955, CLIP-T: 0.9992, VQA-a: 54.6491, VQA-t: 49.1951, Flow-score: 6.6526)

SpargeAttn (CLIPSIM: 0.2044, CLIP-T: 0.9991, VQA-a: 53.2634, VQA-t: 48.8170, Flow-score: 6.1444)

DuSA (CLIPSIM: 0.2065, CLIP-T: 0.9982, VQA-a: 53.5948, VQA-t: 49.1732, Flow-score: 6.7921)

Fig. 6: DuSA achieves noninferior video generation quality compared to VSA (FA2) and outperforms SpargeAttn [4].

➢ The Application of DuSA in Accelerating the Prefilling Stage of LLMs

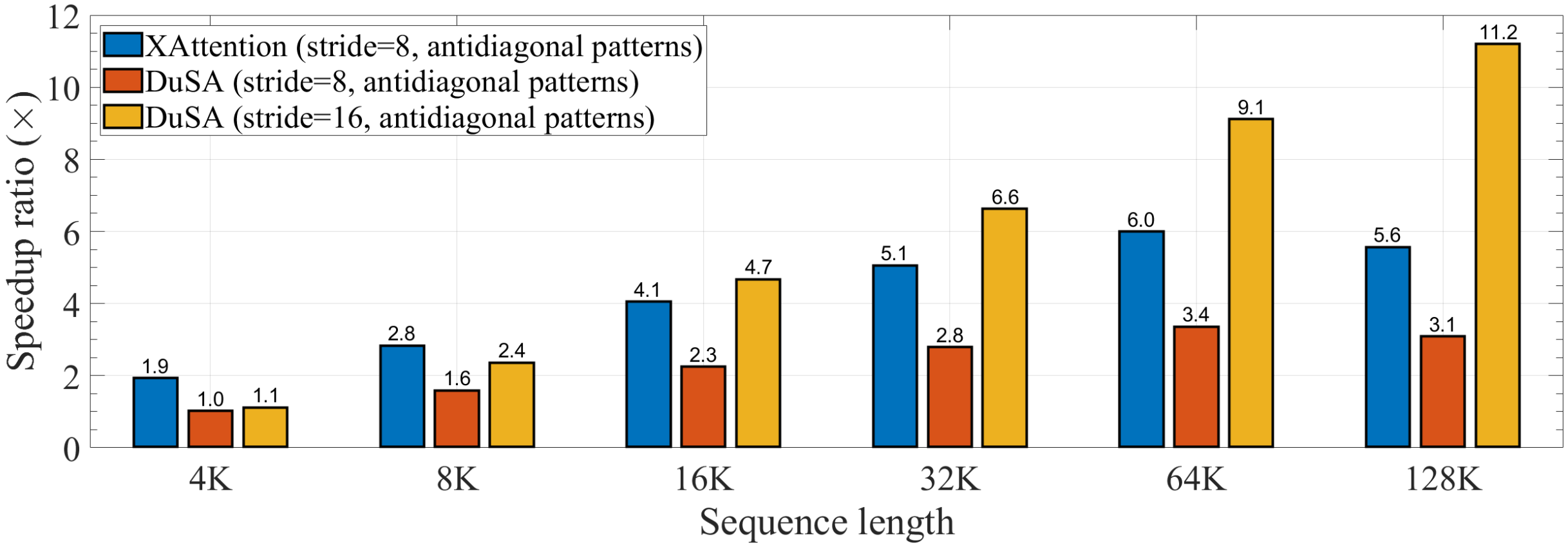
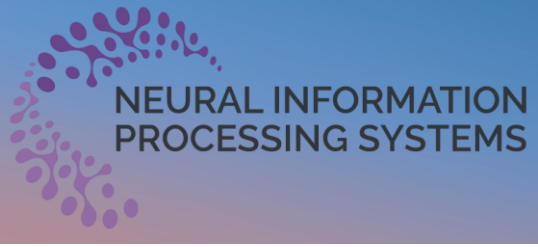


Fig. 7: Speedup relative to FlashAttention of FlashInfer [5] obtained by different methods across different sequence lengths using Llama-3.1-8B-Instruct on 1 NVIDIA H20. DuSA (stride=16) achieve an average score of **41.13** which is closer to the average score of **41.18** of FlashAttention of FlashInfer than the average score of **40.68** of XAttention [6] on LongBench.

➢ Conclusions and References



- A novel and effective dual-stage sparse attention mechanism.
- A mathematical relationship between DuSA and VSA is given.
- DuSA can approximate VSA with a low error.
- DuSA surpasses state-of-the-art methods in different tasks with a more significant acceleration effect.

References:

[1] C. Wu, M. Che, R. Xu, Z. Ran, and H. Yan, "ELFATT: Efficient linear fast attention for vision transformers," in *Proceedings of the 33rd ACM International Conference on Multimedia*, ser. MM '25. New York, NY, USA: Association for Computing Machinery, 2025. [Online]. Available: https://doi.org/10.1145/3746027.3754825

[2] T. Dao, "FlashAttention-2: Faster attention with better parallelism and work partitioning," in *International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=mZn2Xyh9Ec

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017, pp. 5998–6008.

[4] J. Zhang, C. Xiang, H. Huang, J. Wei, H. Xi, J. Zhu, and J. Chen, "SpargeAttention: Accurate and training-free sparse attention accelerating any model inference," in *Proceedings of the 42nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Singh, M. Fazel, D. Hsu, S. Lacoste-Julien, F. Berkenkamp, T. Maharaj, K. Wagstaff, and J. Zhu, Eds., vol. 267. PMLR, 2025, pp. 76397–76413.

[5] Z. Ye, R. Lai, B.-R. Lu, C.-Y. Lin, S. Zheng, L. Chen, T. Chen, and L. Ceze, "Cascade inference: Memory bandwidth efficient shared prefix batch decoding," February 2024. [Online]. Available: https://flashinfer.ai/2024/02/02/cascade-inference.html

[6] R. Xu, G. Xiao, H. Huang, J. Guo, and S. Han, "XAttention: Block sparse attention with antidiagonal scoring," in *Proceedings of the 42nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Singh, M. Fazel, D. Hsu, S. Lacoste-Julien, F. Berkenkamp, T. Maharaj, K. Wagstaff, and J. Zhu, Eds., vol. 267. PMLR, 2025, pp. 69819–69831.