Table 1: **Inference speed comparison of SpikingBrain-7B under sequence parallelism (SP).** The metric is TTFT (ms), defined as the latency to complete prefill and generate the first token for a given prompt length. In SP configuration, our 7B model uses ZeCO for the linear attention module and P2P communication for the SWA module, while the Qwen2.5 baseline employs A2A communication. All measurements are conducted on NVIDIA H100 GPUs and averaged over 10 runs. "–" indicates infeasible measurement due to resource constraints.

Sequence length	GPU count	**SpikingBrain-7B**	Qwen2.5-7B
256k	8	1015	7419
512k	16	1037	14398
1M	32	1054	27929
2M	64	1070	–
4M	128	1073	–