Rebuttal Material for ICML 2025

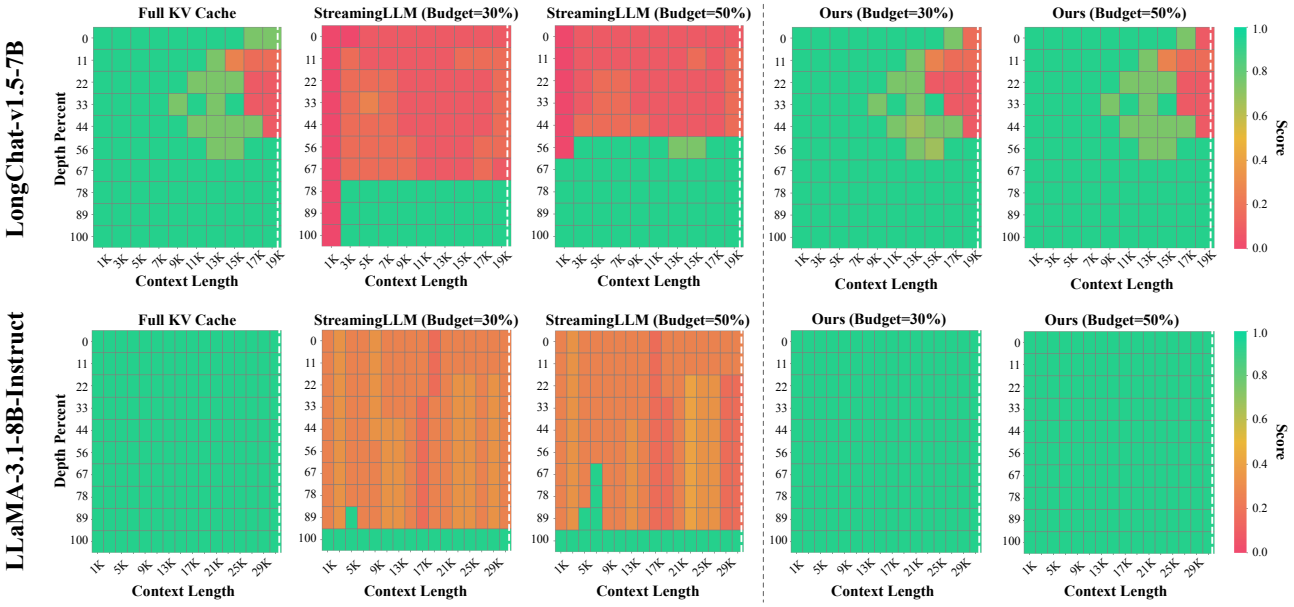

Figure 1. Needle-in-a-Haystack benchmark comparison. The x-axis denotes the length of the context ("haystack") and the y-axis indicates the position where the "needle" (a short prompt) is inserted within the context. We set the context length to 0.5K-20K on LongChat-v1.5-7B and 1K-32K on LLaMA-3.1-8B-Instruct. Our method demonstrates comparable performance with the full KV cache and outstanding performance compared with StreamingLLM.

Table 1. Extend the method of MPCache to LLaMA-3.1-8B-Instruct on LongBench with an average KV cache budget of 2048.

Method	NarrativeQA	Qasper	HotpotQA	2WikiMQA	MuSiQue	TREC	SAMSum	TriviaQA	QMSum	PR-en	MF-en
Full Cache	30.21	45.52	55.53	46.71	31.34	72.50	43.86	91.74	25.20	99.50	54.94
StreamingLLM	26.64	30.77	49.23	44.66	24.31	67.50	42.49	90.98	21.67	87.00	37.85
DuoAttention	25.61	42.31	52.36	42.14	28.17	66.00	43.36	89.93	22.11	<u>98.50</u>	**55.53**
SnapKV	<u>28.53</u>	39.13	<u>54.32</u>	<u>46.59</u>	28.48	55.55	43.10	**92.04**	23.92	**99.50**	53.91
Quest	28.00	<u>45.55</u>	53.73	44.76	<u>29.82</u>	<u>68.56</u>	**44.05**	90.90	**24.91**	**99.50**	<u>55.40</u>
MPCache	**30.17**	**46.11**	**55.21**	**46.61**	**30.49**	**69.50**	<u>43.67</u>	<u>91.53</u>	<u>24.83</u>	**99.50**	53.41