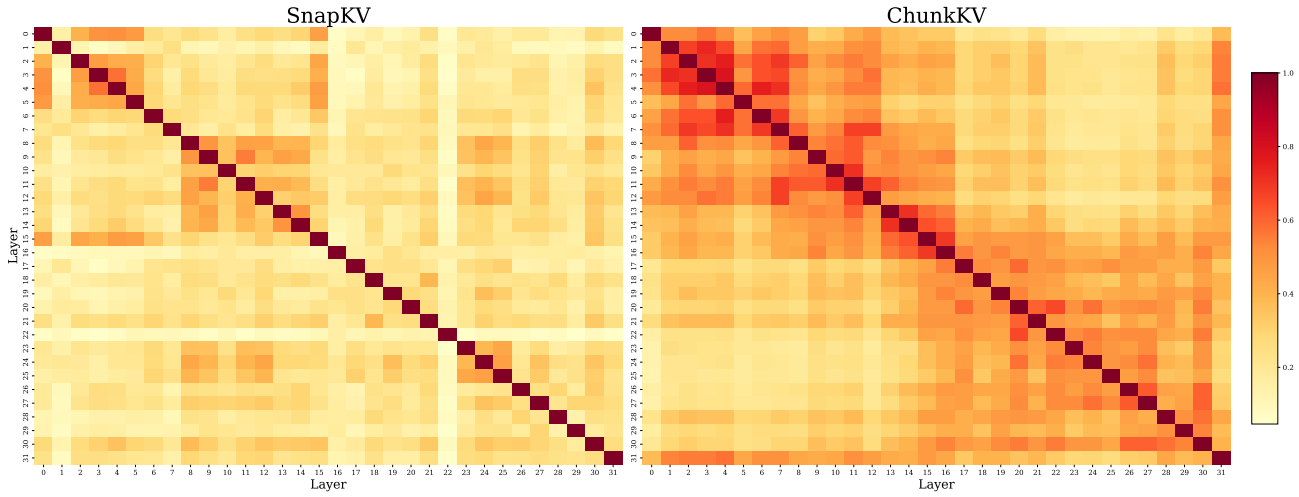

Figure 2: Layer-wise similarity heatmaps of the preserved KV cache indices by SnapKV (left) and ChunkKV (right) on LLaMA-3-8B-Instruct.

continuous sequence in ChunkKV is better than according to sparse tokens. Informally speaking, the continuously chunk-level KV cache preserves the whole examples (semantic information) in ICL, thus reducing the requirement on distinguishability, i.e lower bound of KL divergence between the example and the question (Equation 4 in Condition 2). The complete analysis is provided in Appendix C.

4. Experiment Results

In this section, we conduct experiments to evaluate the effectiveness of ChunkKV on KV cache compression in two benchmark fields, with a chunk size set to 10 even for various model architectures. The first is the In-Context Learning benchmark, for which we select GSM8K (Cobbe et al., 2021) and Jailbreakv (Luo et al., 2024) to evaluate the performance of ChunkKV, furthermore we also include multi-step reasoning LLM DeepSeek-R1-Distill-Llama-8B (Guo et al., 2025) to evaluate the performance of ChunkKV. The In-Context Learning scenario is a crucial capability for LLMs and has been adapted in many powerful technologies such as Chain-of-Thought (Wei et al., 2022; Diao et al., 2024; Pan et al., 2024b). The second is the Long-Context benchmark, which includes LongBench (Bai et al., 2024) and Needle-In-A-HayStack (NIAH) (Kamradt, 2023), both widely used for assessing KV cache compression methods. All experiments were conducted three times, using the mean score to ensure robustness.

4.1. In-Context Learning

The In-Context Learning (ICL) ability significantly enhances the impact of prompts on large language models (LLMs). For example, the Chain-of-Thought approach (Wei et al., 2022) increases the accuracy of the GSM8K of the PaLM model (Chowdhery et al., 2022) from 18% to 57% without additional training. In this section, we evaluate the performance of ChunkKV on the GSM8K, Many-Shot GSM8K (Agarwal et al., 2024), and JailbreakV (Luo et al., 2024) benchmarks.

Table 3: GSM8K Performance Comparison.

Ratio	StreamingLLM	H2O	SnapKV	PyramidKV	**ChunkKV (Ours)**
DeepSeek-R1-Distill-Llama-8B FullKV: 69.4% ↑					
10%	51.6%	55.6%	57.6%	62.6%	**65.7%**
LlaMa-3.1-8B-Instruct FullKV: 79.5% ↑					
30%	70.5%	72.2%	76.1%	77.1%	**77.3%**
20%	63.8%	64.0%	68.8%	71.4%	**77.6%**
10%	47.8%	45.0%	50.3%	48.2%	**65.7%**
LlaMa-3-8B-Instruct FullKV: 76.8% ↑					
30%	70.6%	73.6%	70.2%	68.2%	**74.6%**
Qwen2-7B-Instruct FullKV: 71.1% ↑					
30%	70.8%	61.2%	70.8%	64.7%	**73.5%**

GSM8K In the in-context learning scenario, we evaluated multiple KV cache compression methods for GSM8K (Cobbe et al., 2021), which contains more than 1,000 arithmetic questions on LLaMA-3-8B-Instruct, LLaMA-3.1-8B-Instruct (Meta, 2024), Qwen2-7B-Instruct (Yang et al., 2024a) and DeepSeek-R1-Distill-Llama-8B (Guo et al., 2025). Follow the Agarwal et al. (2024), we consider many-shot GSM8K as a long-context reasoning scenario, which is a more challenging task than LongBench (Bai et al., 2024). The CoT prompt settings for this experiment are the same as those used by Wei et al. (2022), for many-shot GSM8K we set the number of shots to 50, which the prompt length is more than 4k tokens.