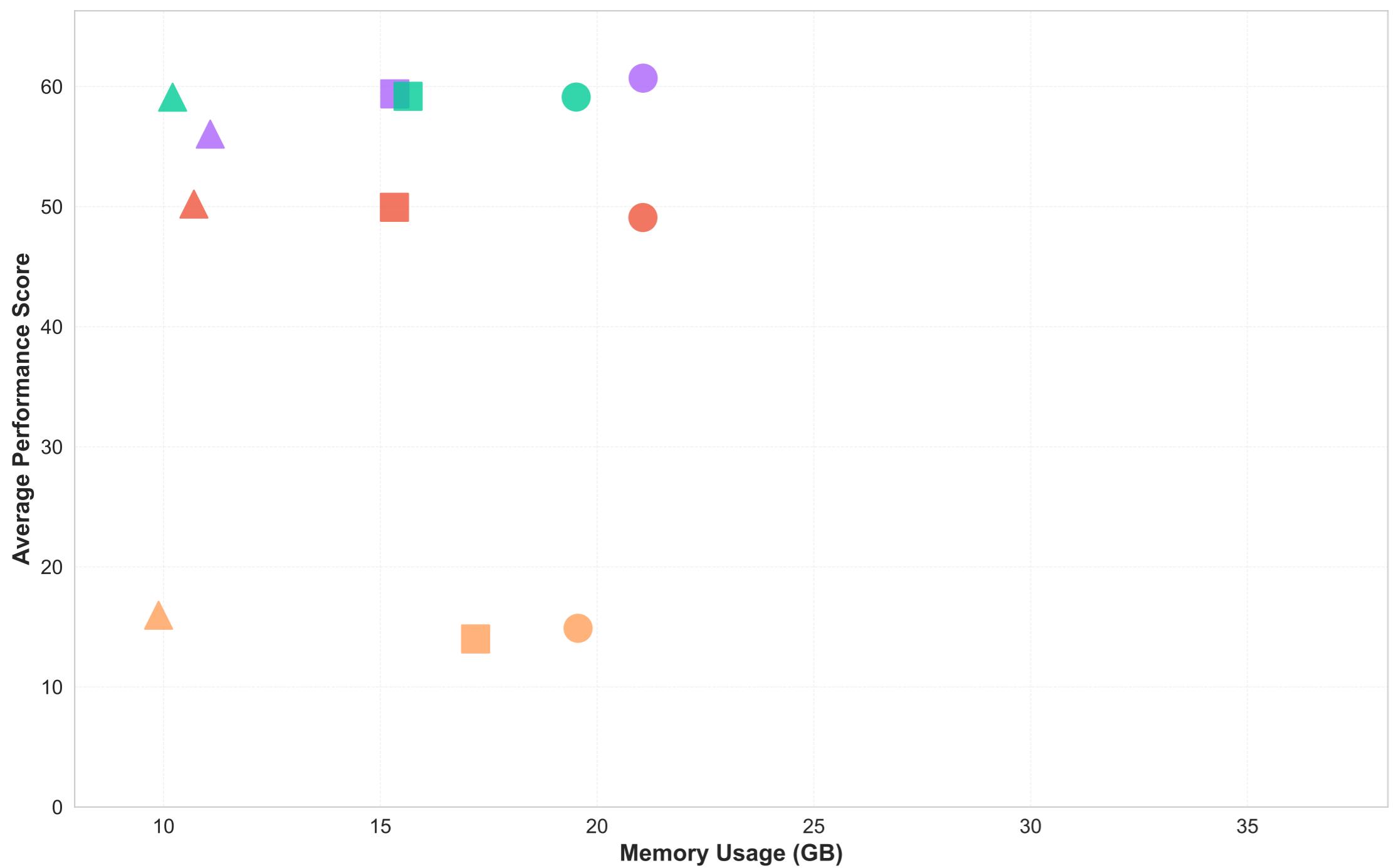
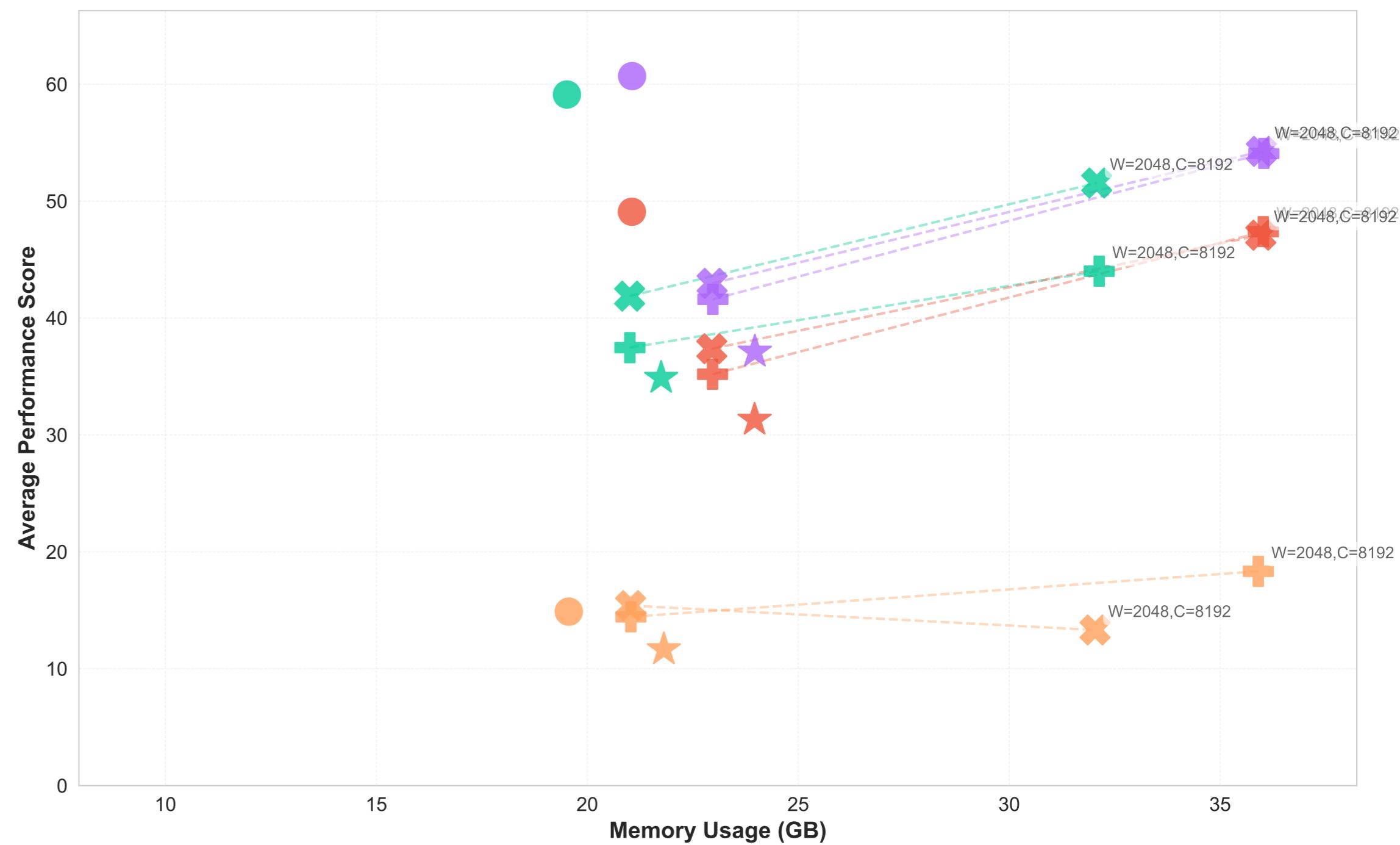


Model Comparison: Average Performance vs Memory Usage Combined HELMET (16K) and LongProc (2K) Benchmarks

Quantization Methods



KV Cache Methods



Models

● Llama-3.1-8B-Instruct ● Qwen2.5-7B-Instruct ● DeepSeek-R1-Distill-Llama-8B ● DeepSeek-R1-Distill-Qwen-7B

Techniques

▲ NF4 ■ Int8 ● Baseline + PyramidKV ✕ SnapKV ★ StreamingLLM