A. In-depth Analysis of ChunkKV vs. Discrete Token Methods

A.1. Quantitative Analysis

To rigorously evaluate the effectiveness of ChunkKV compared to discrete token-based methods, we conducted systematic experiments using a LLaMA-3-8B-Instruct model. We randomly selected 100 sequences from the each sub-category of LongBench dataset and analyzed two key metrics across different model layers: KV cache L1 loss and attention cosine similarity. For each sequence, we: 1. Computed the full KV cache and attention patterns without compression as ground truth. 2. Applied ChunkKV, SnapKV, and H2O compression methods with a fixed 10% compression ratio, and the parameters of the three methods are set the same as in Table 14. 3. Measured the differences between compressed and uncompressed versions.

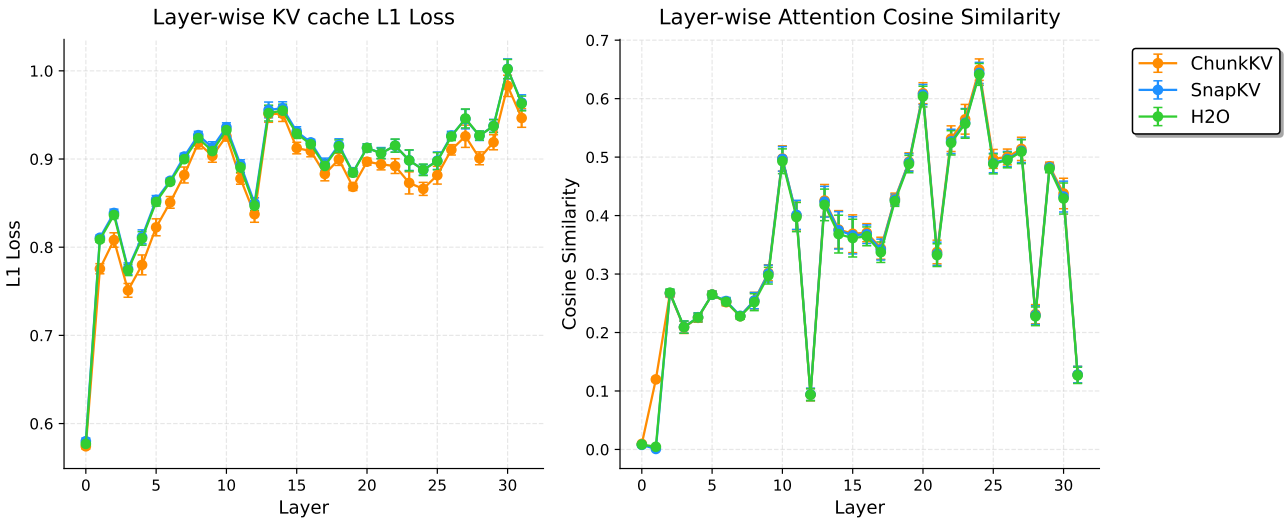

Figure 6: Layer-wise comparison of L1 loss and attention cosine similarity between ChunkKV and discrete token-based methods in Single-Document QA sub-category of LongBench.

Results Analysis As shown in Figure 6, ChunkKV demonstrates superior performance across both metrics:

- **KV Cache L1 Loss:** ChunkKV achieves consistently lower L1 loss compared to SnapKV and H2O, particularly in the early and middle layers (layers 5-25). This indicates better preservation of the original KV cache information through the semantic chunk-based approach.
- **Attention Cosine Similarity:** ChunkKV exhibits higher similarity scores across most layers, with notably strong performance in layers 0-5 and 20-30. This suggests better preservation of attention relationships between tokens, which is crucial for maintaining semantic understanding.

To quantify these improvements, we calculated average metrics across all layers, as shown in Table 11. ChunkKV achieves both the lowest L1 loss and highest attention cosine similarity, outperforming both baseline methods.

Significance of Results While the improvements may appear modest in absolute terms (approximately 2% in L1 loss and 1.5% in cosine similarity), their practical significance is substantial. These metrics reflect the model's ability to maintain crucial semantic relationships and attention patterns, which are essential for complex reasoning tasks. The consistent improvements across different sequences demonstrate that preserving semantic chunks leads to better information retention than selecting individual tokens.

The enhanced performance is particularly evident in the middle layers of the model, which are typically responsible for higher-level semantic processing. This provides concrete evidence for why ChunkKV achieves superior performance on downstream tasks compared to discrete token-based methods.