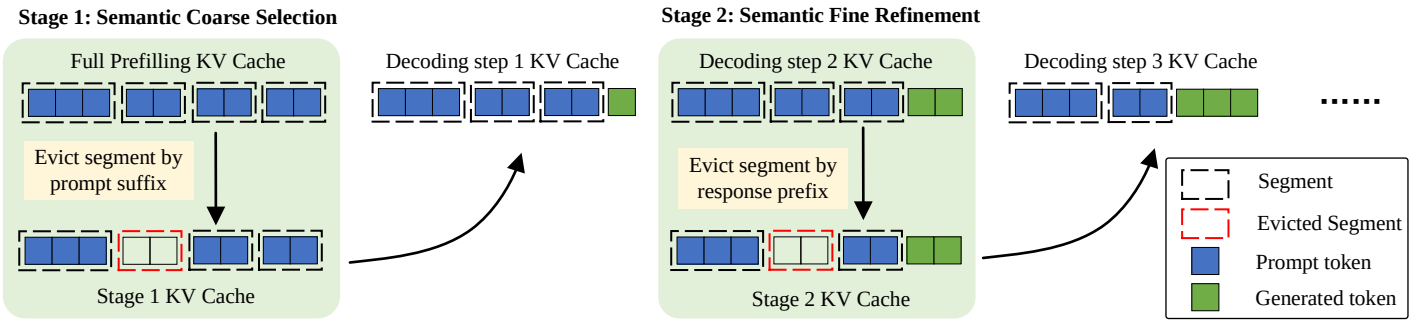
Stage 1: Semantic Coarse Selection
Full Prefilling KV Cache
Evict segment by prompt suffix
Stage 1 KV Cache
Decoding step 1 KV Cache
Stage 2: Semantic Fine Refinement
Decoding step 2 KV Cache
Evict segment by response prefix
Stage 2 KV Cache
Decoding step 3 KV Cache
.....
Segment
Evicted Segment
Prompt token
Generated token