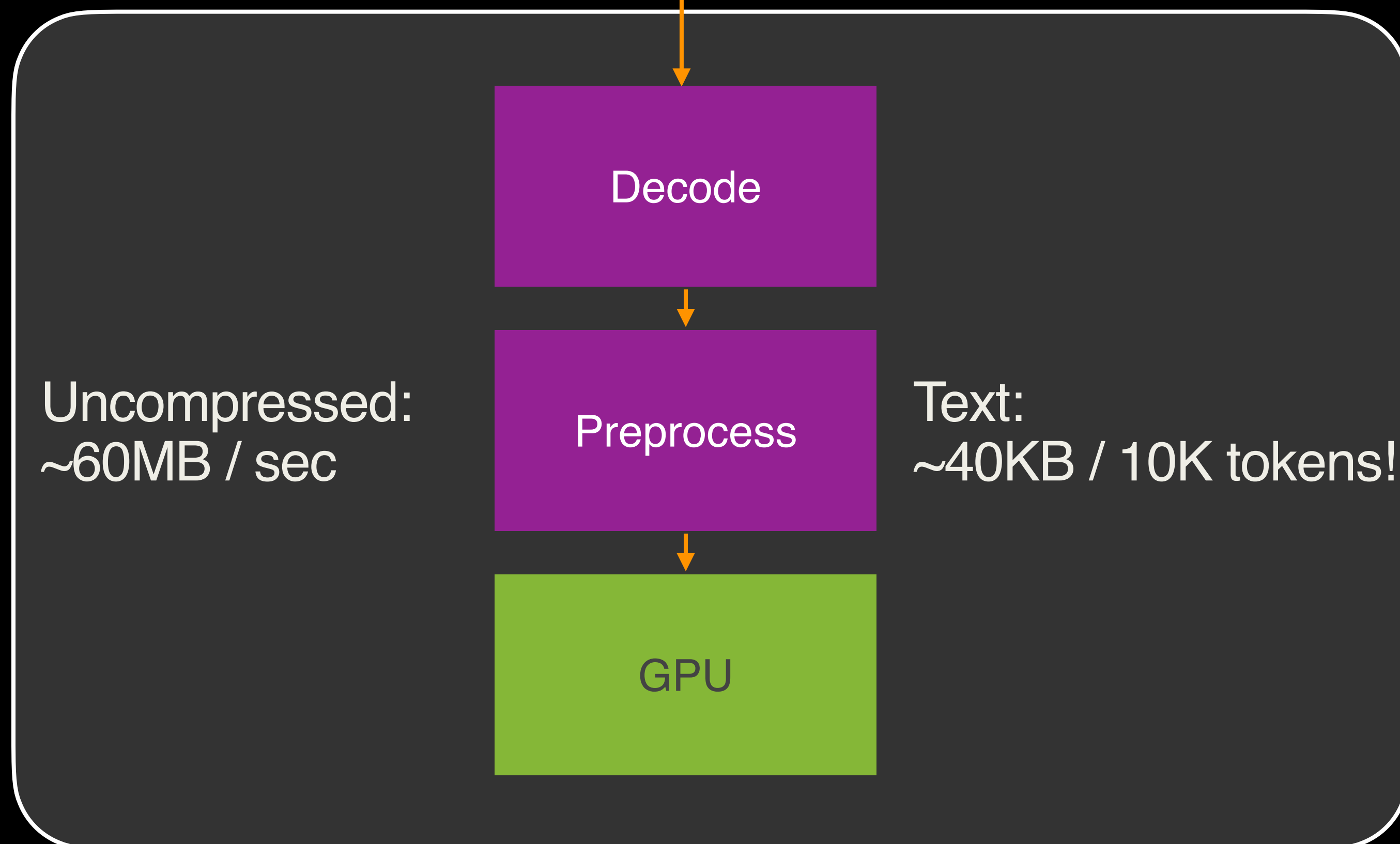




Compressed data size:  
 $O(10\text{KB})$  / 1080p frame  
 $O(1\text{MB})$  / sec





Compressed data size:  
 $O(10\text{KB})$  / 1080p frame  
 $O(1\text{MB})$  / sec

Decode

Preprocess

GPU

Uncompressed:  
 $\sim 60\text{MB}$  / sec

Text:  
 $\sim 40\text{KB}$  / 10K tokens!

Don't Bottleneck the GPU!