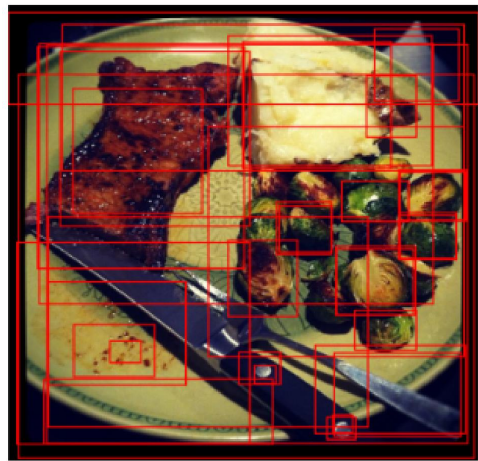
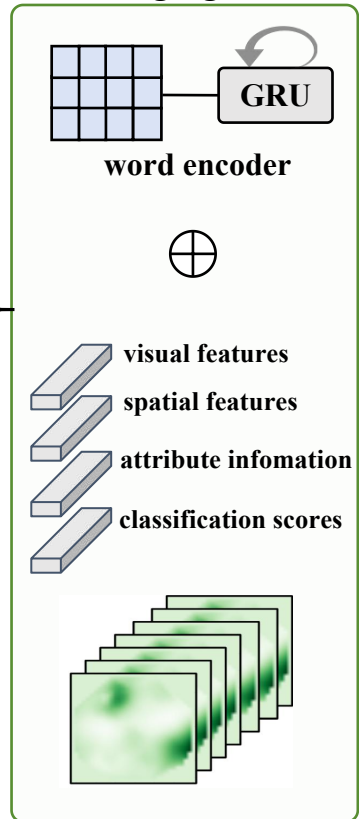


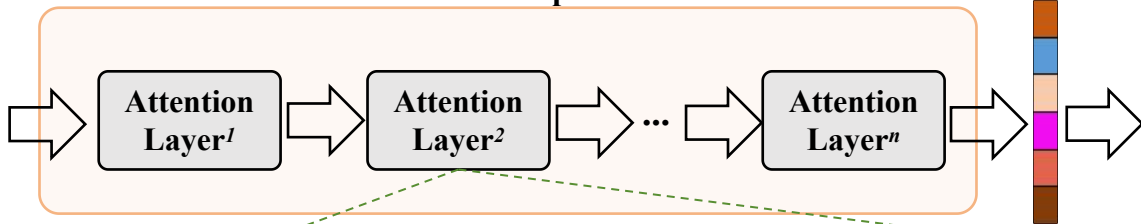
Question: What color are the vegetables on the plate?



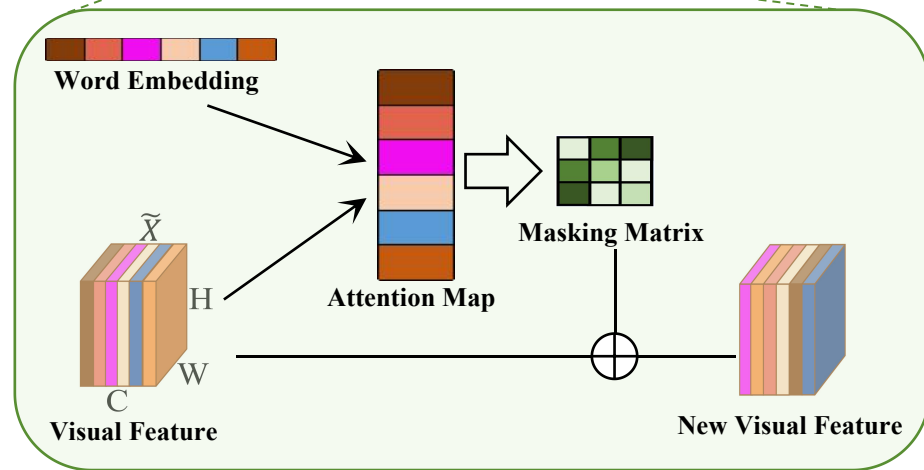
Visual Encoder: extracting visual-language features



Cross-Modal Retrieval: multi-round retrieval information process



Answer: Green.



Mask irrelevant features