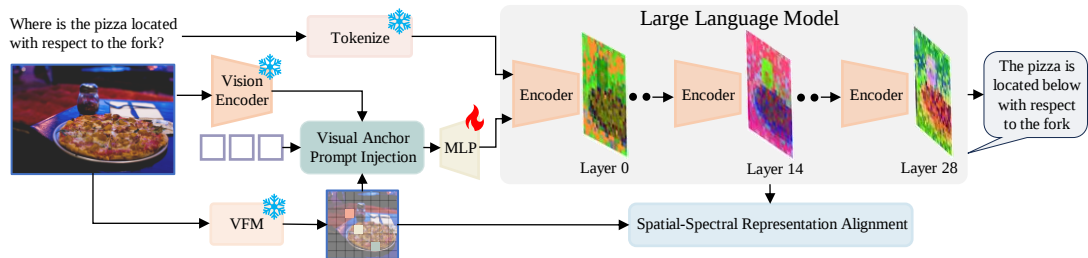
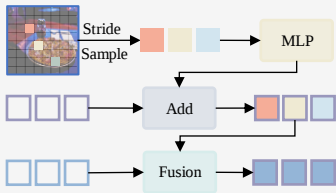


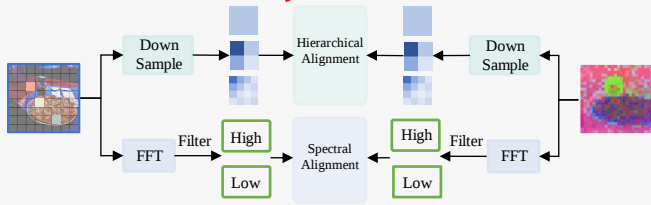
Where is the pizza located with respect to the fork?



Visual Anchor Prompt Injection 



Spatial-Spectral Representation Alignment 



VFM Vision Foundation Model  Vision Prompt  Vision Feature  Enhanced Feature  Fusion  Multi-Head Cross Attention  Trainable  Frozen