

  Action Token   Visual Token  Language Token $\rightarrow$ Cross Attention

