Ob man encoder-decoder nimmt, also mask-then-predict, oder nur einen encoder mit student-teacher ändert nicht groß was an der größe, da in beiden Fällen „2“ modelle genommen werden