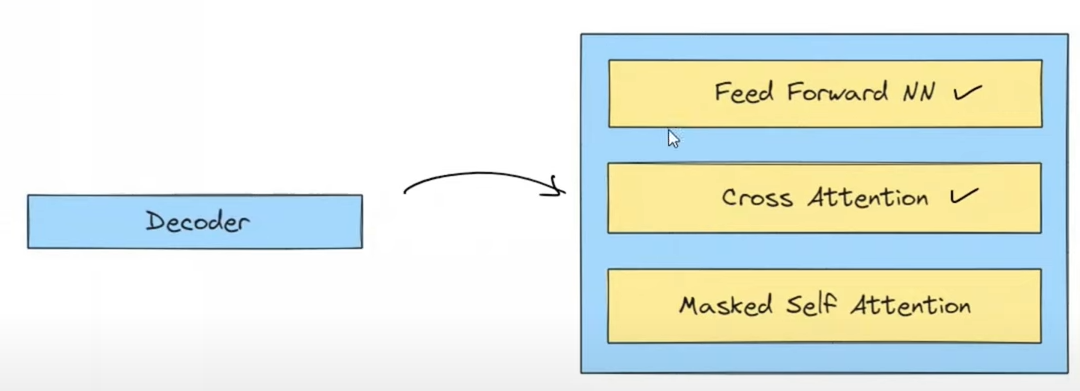
## horizontal line

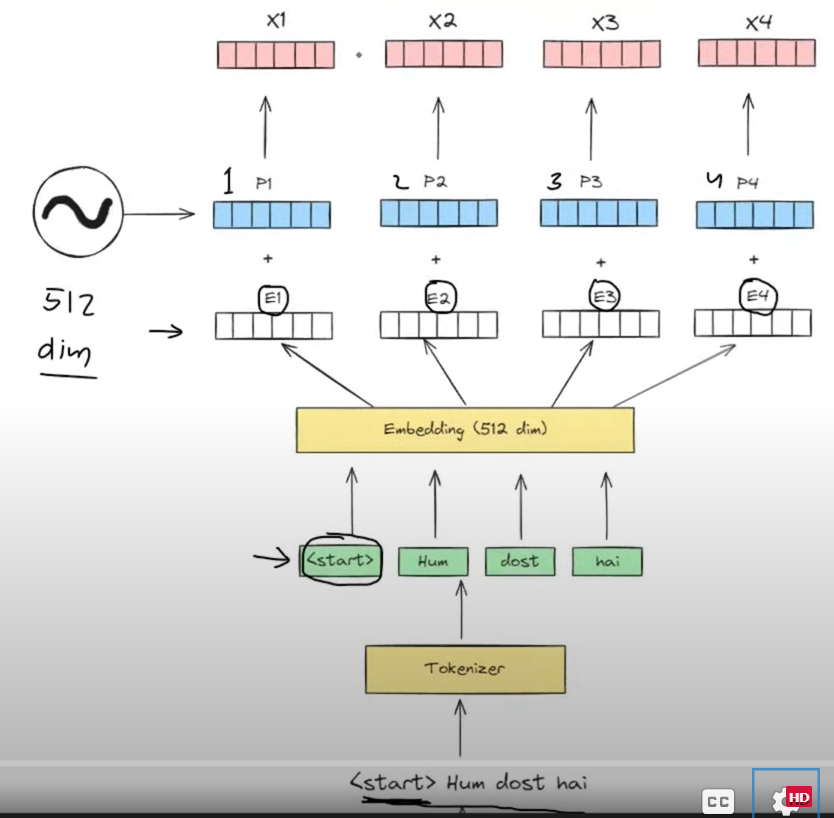
Transformer Decoder

12.06.2025

# Transformer Decoder during Training

Decoder also composed of 6 decoder blocks that are identical but with varying parameters in each block. 

Encoder is same at both aspects but training is non AR and prediction is AR in decoder.



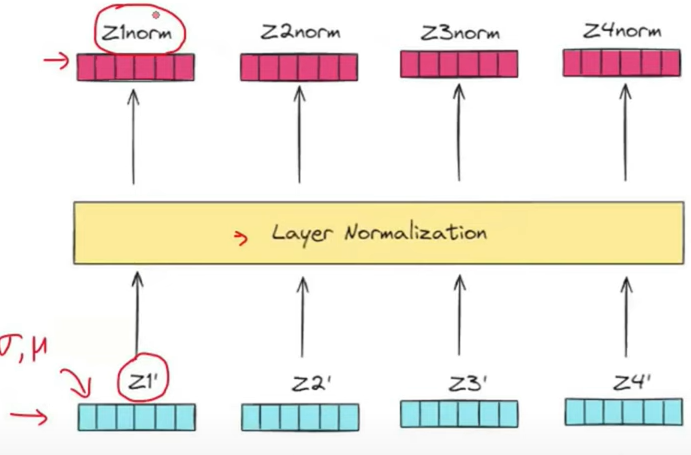
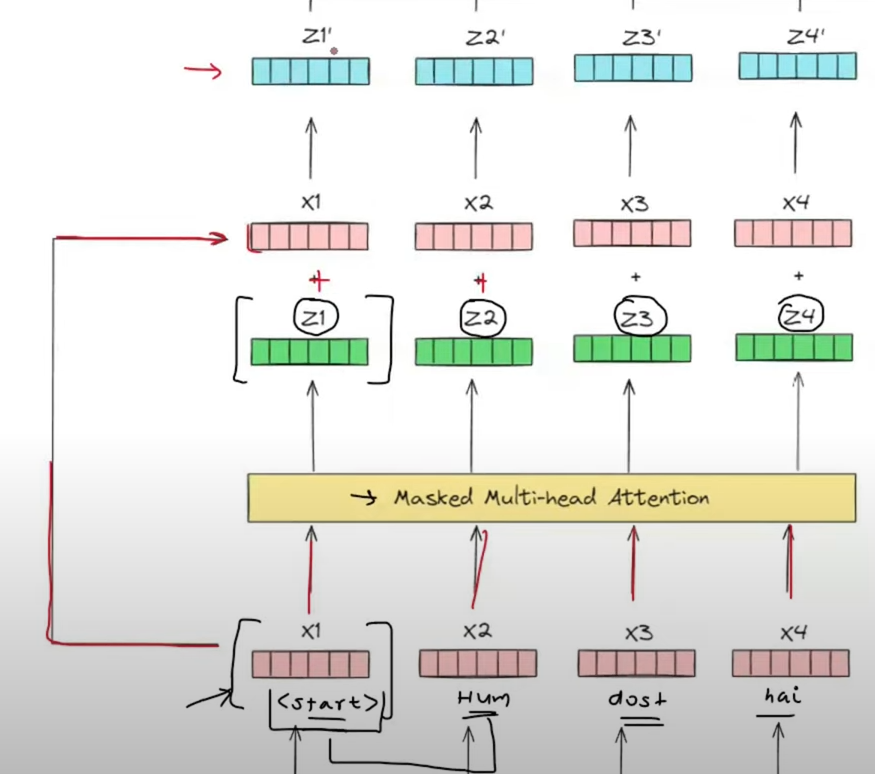
Now we see first decoder in training

In Input Block going on 4 processes :

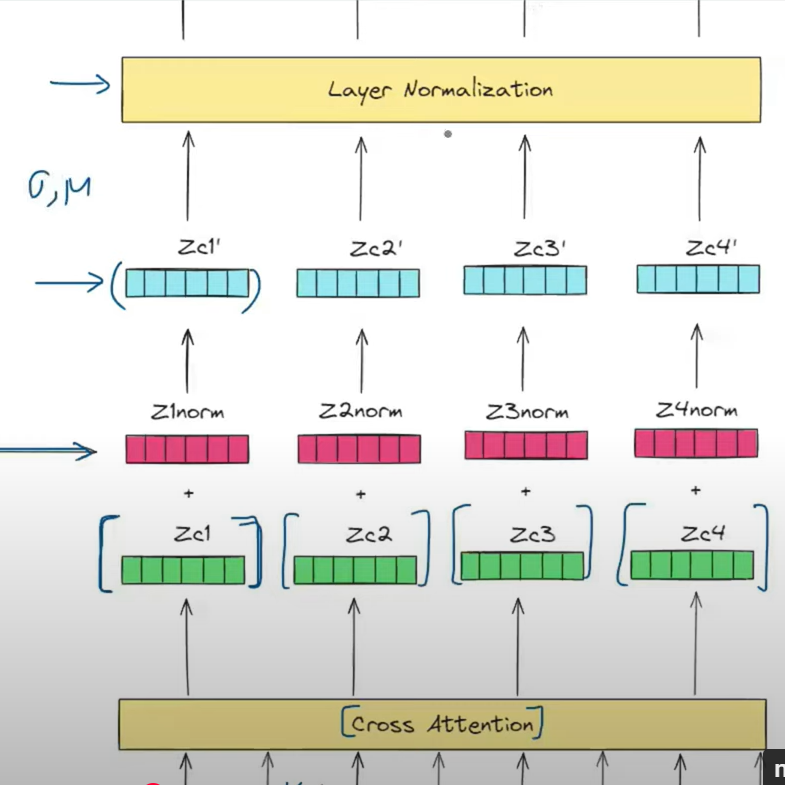
1. Shifting (Right shift)
2. Tokenization
3. Embedding
4. Positional Encoding

E.g. We are Friends (हम दोस्त है)

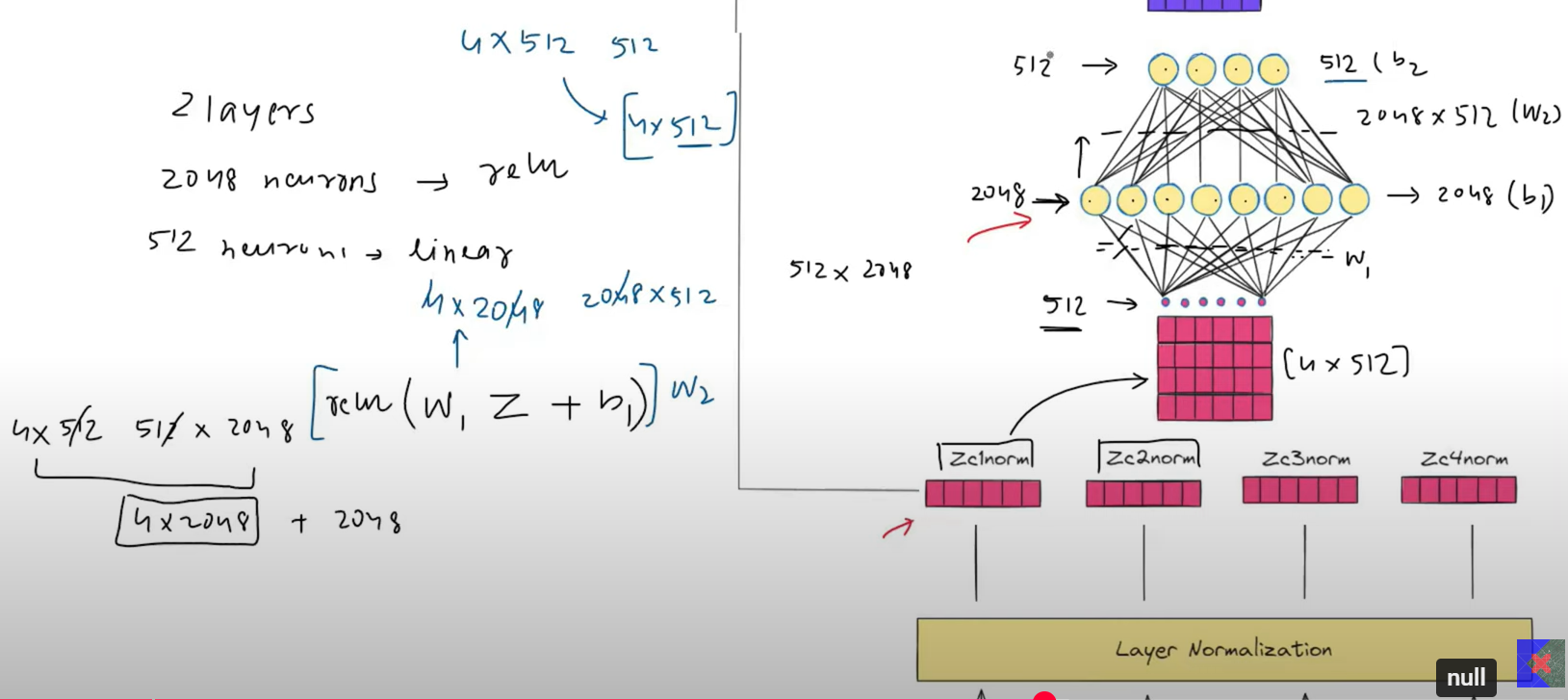
## Now first we have Masked Multi Head Attention Block

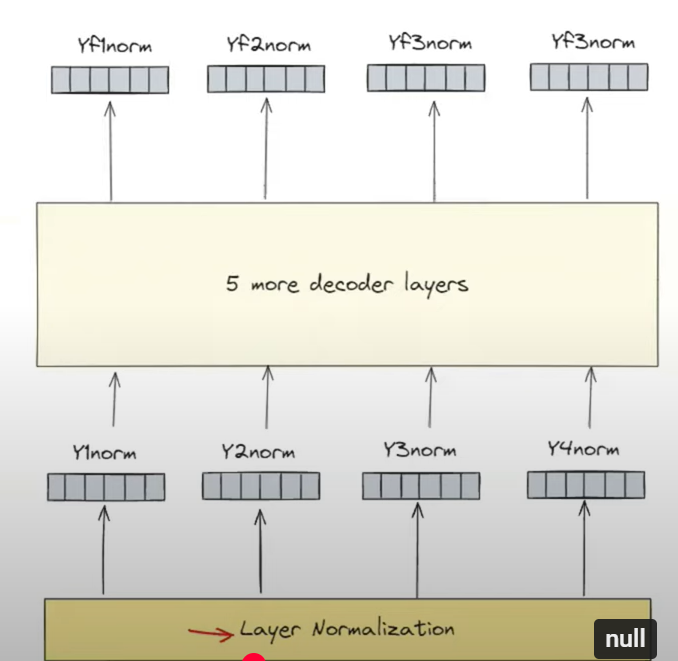
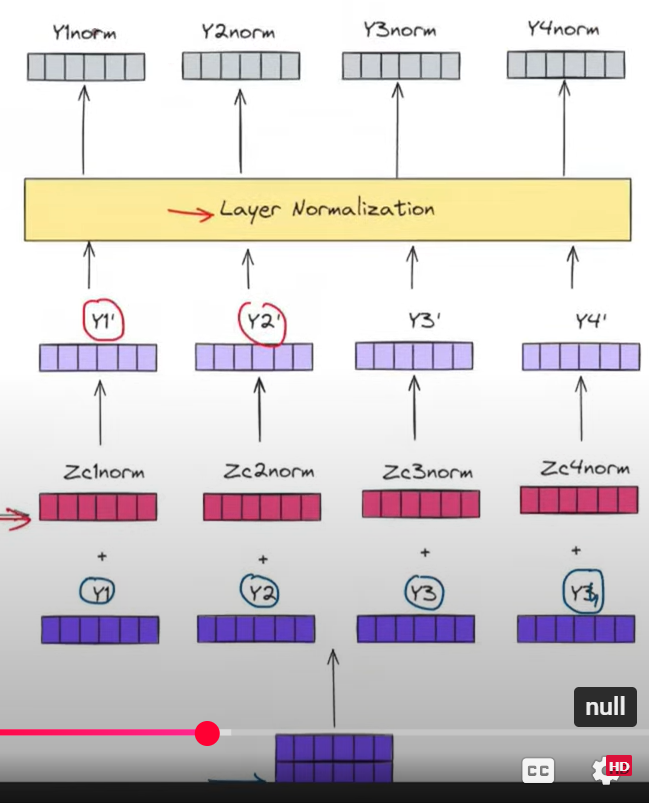
Addition and Layer Normalization is also applied after every operation because after multiple operations may be the vector has unnormalized values that need to be scaled in.

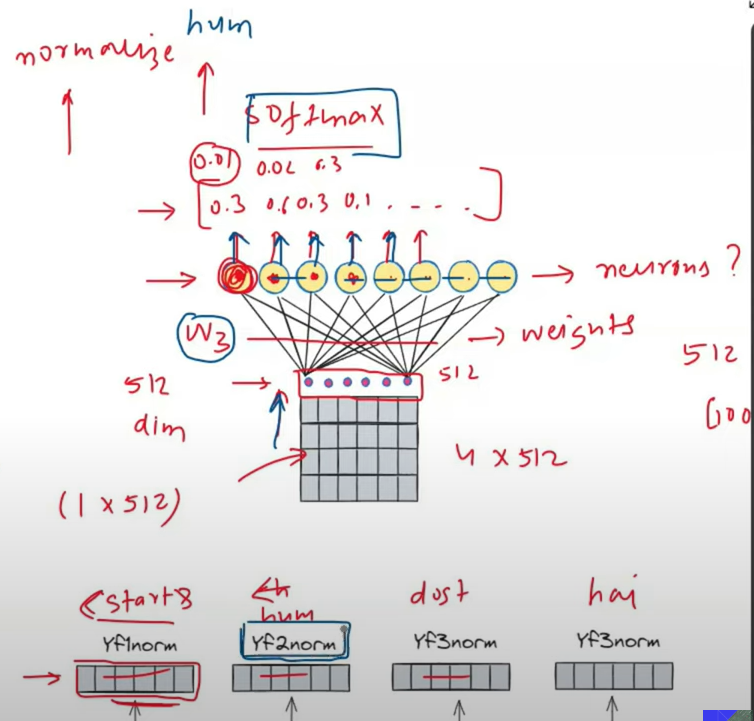
## 2nd block is of Cross Attention



## 3rd block is Feed Forward Neural Network







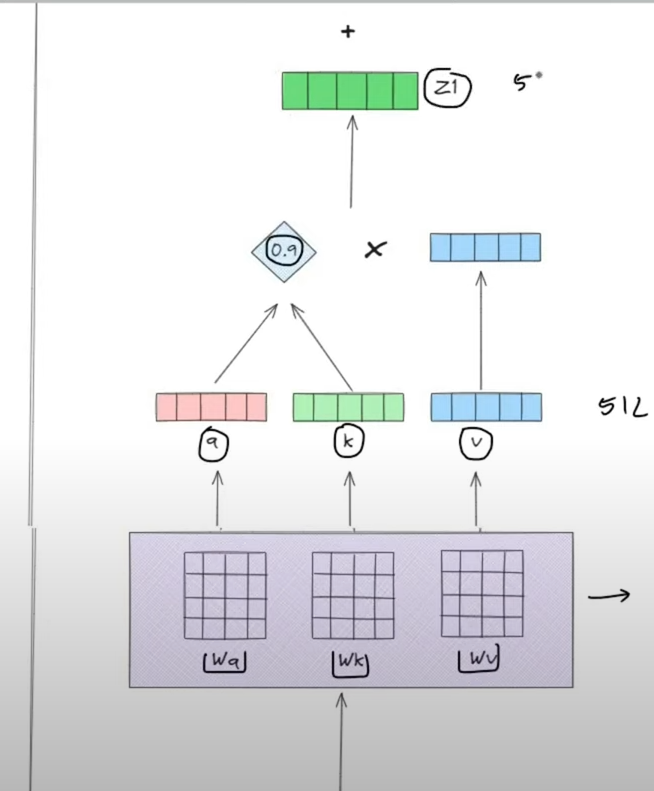
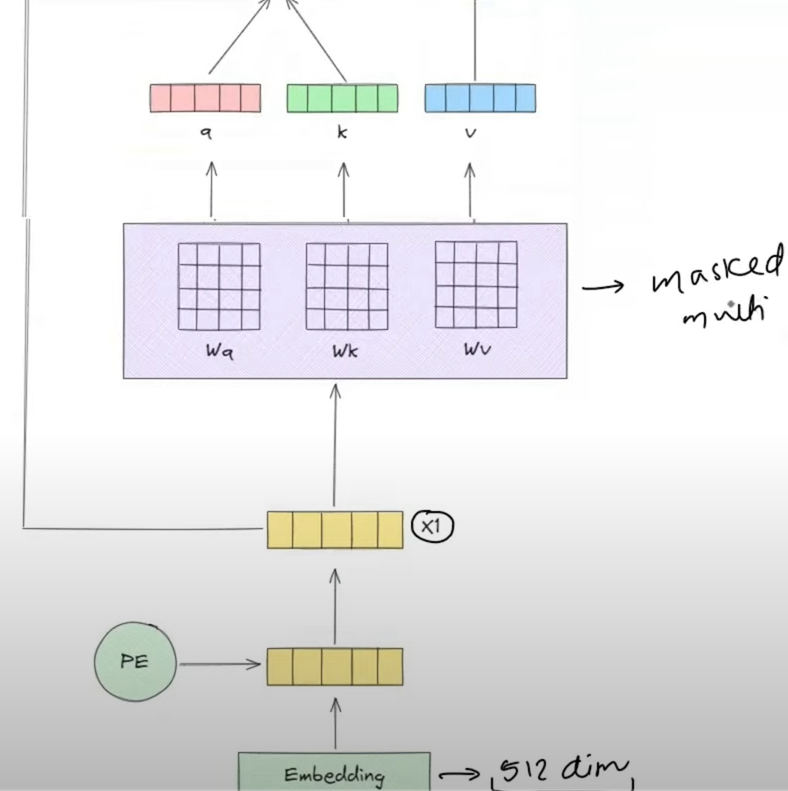
Output :

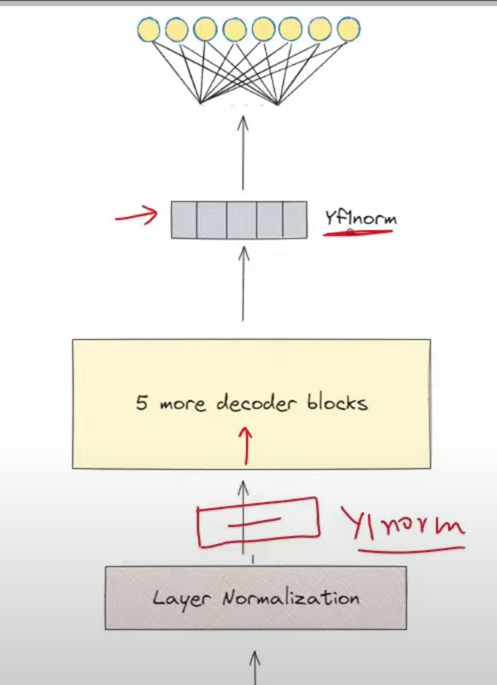
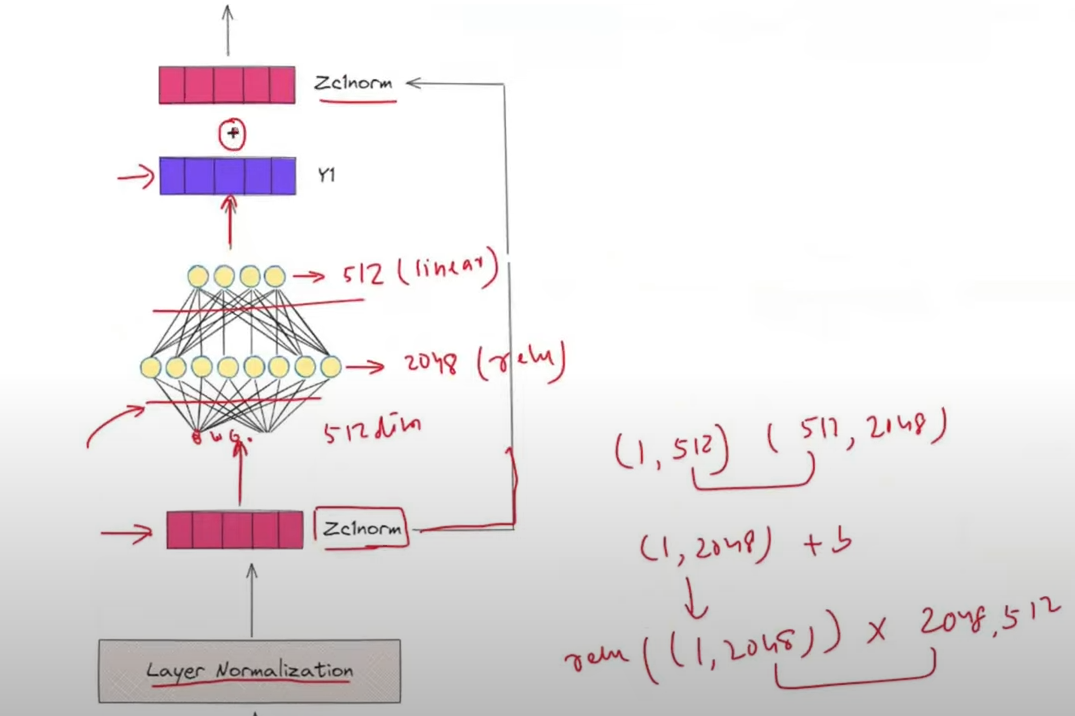
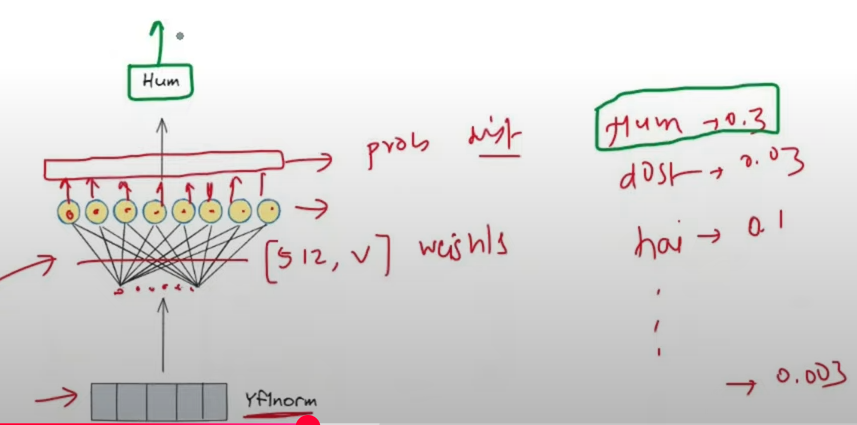
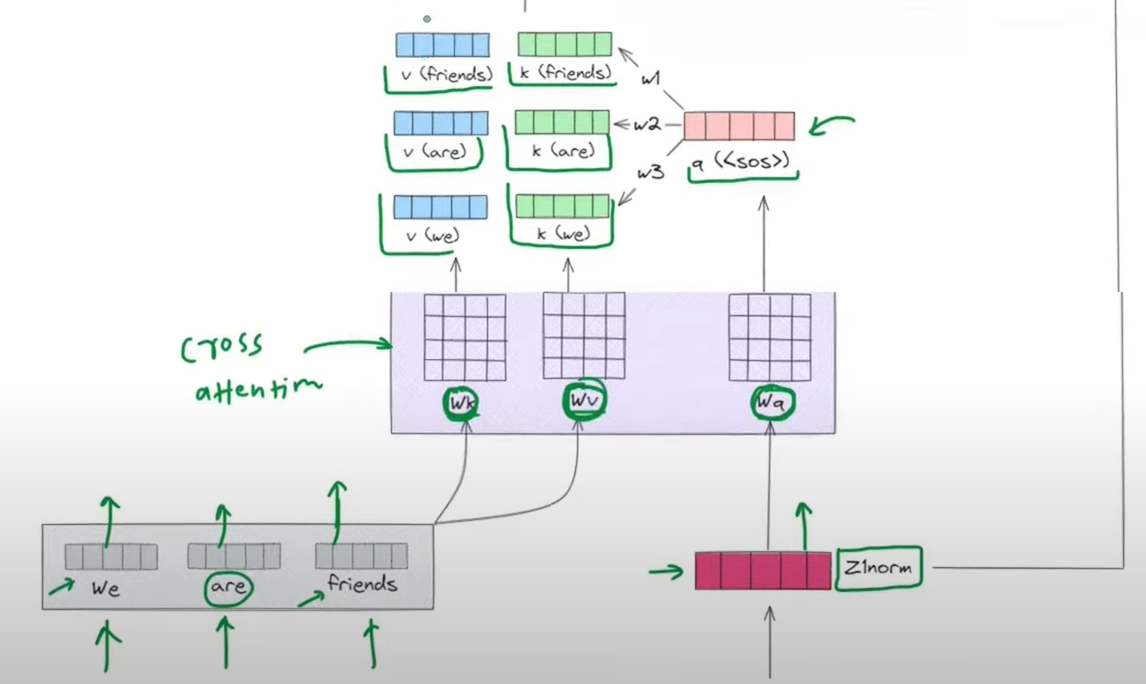
At output no of nodes is same as all the words in vocab of output sequence. We calculate probability of all the unique words in vocab at output sequence.

The word with max probability is the final output.

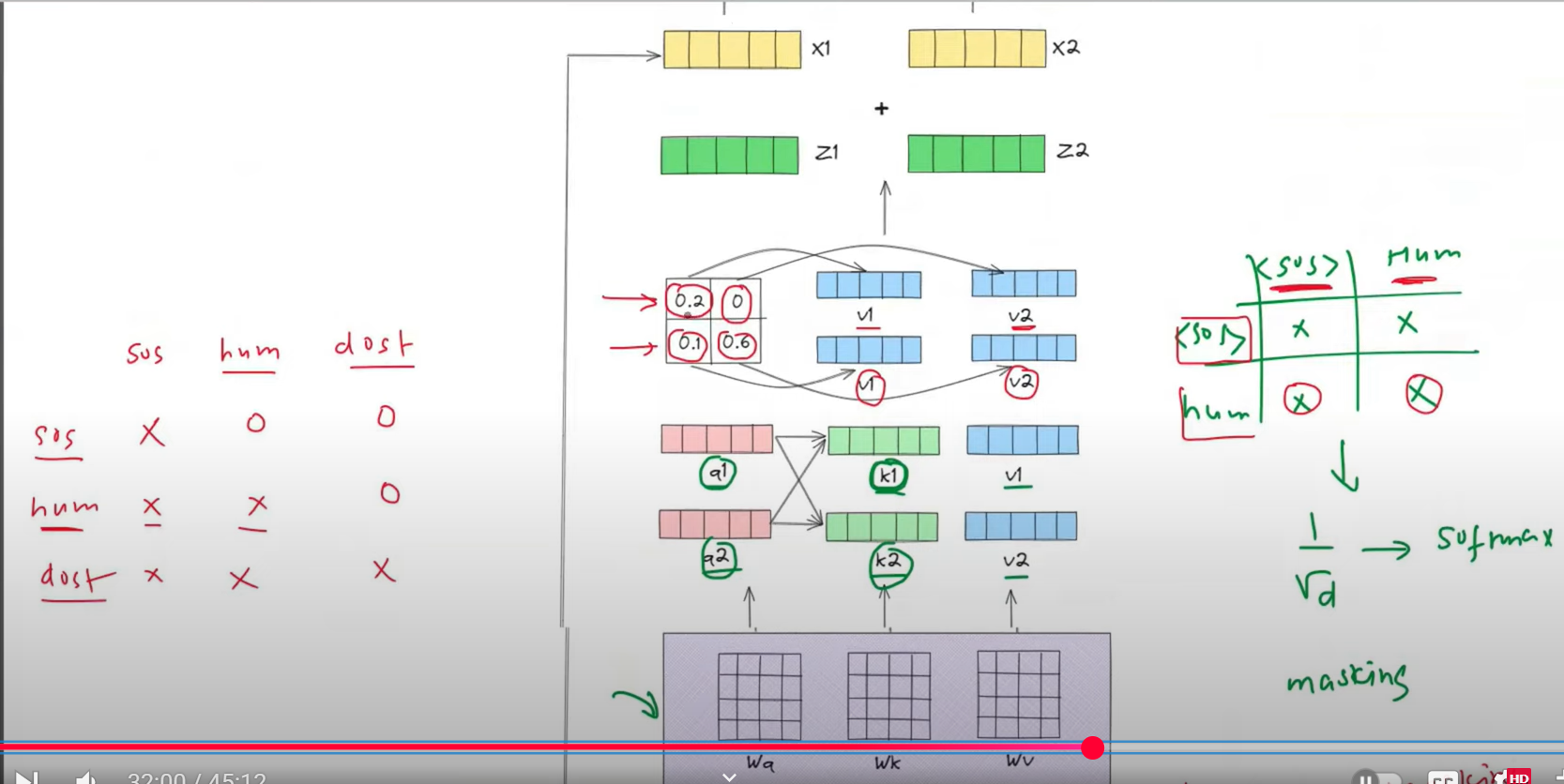
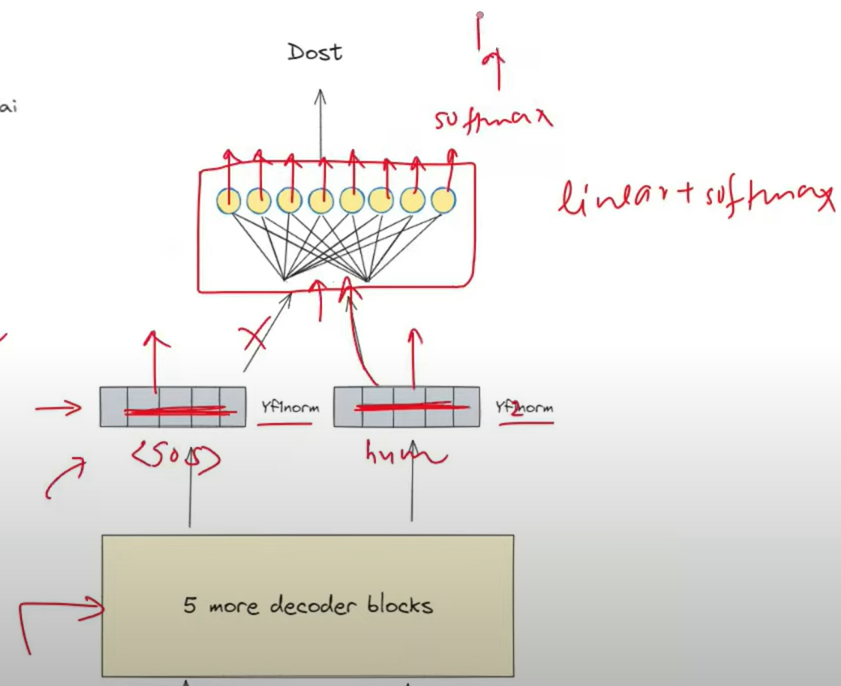
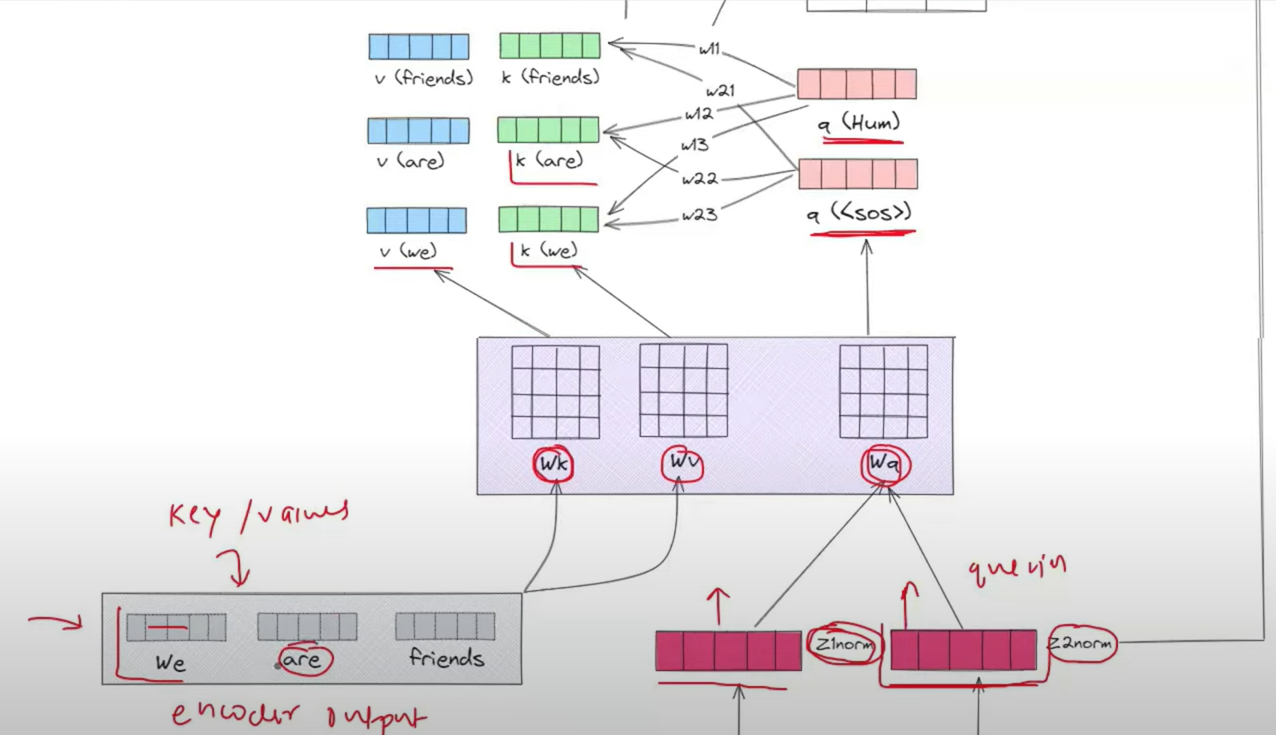
# Transformer decoder during Interference

See what happens at t=1 when we give <sos>.





Similarly : when 2 tokens were sent the main thing is that based on the latest token output is generated and similarly goes for the next ones.



Either we dive 2 tokens or more only the last one is sent to process next one i.e the latest one.