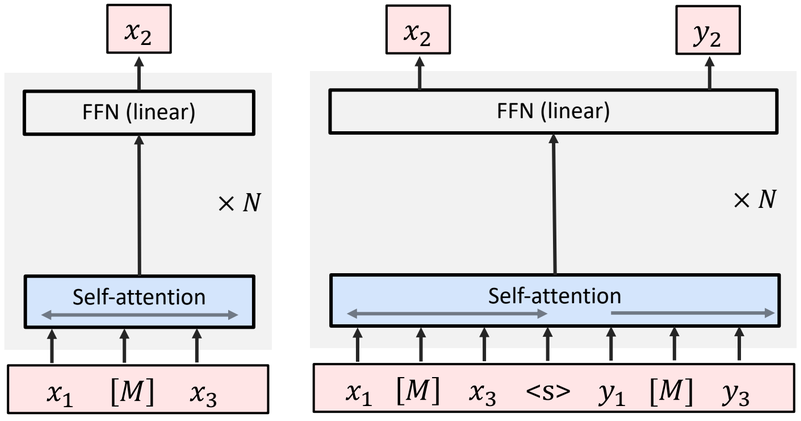
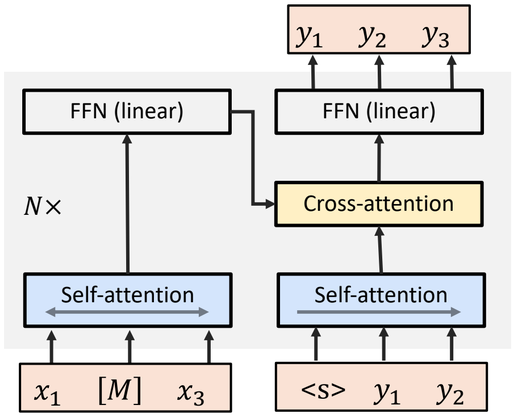
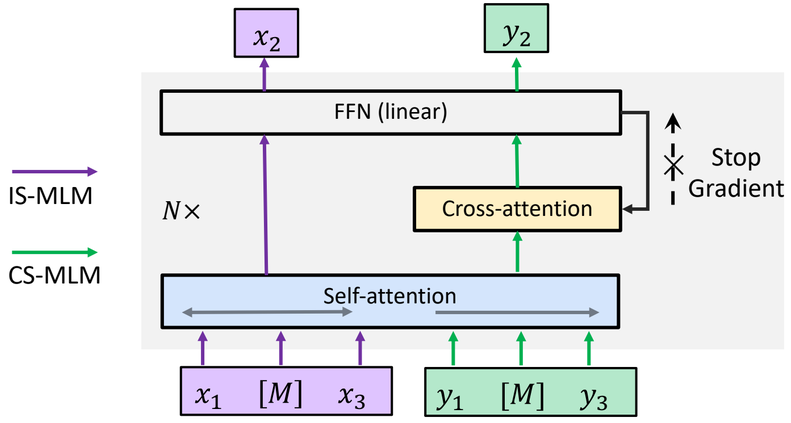


(a) Encoder-Only Models (e.g., Left: BERT, XLM-R; Right: UNILM)



(b) Encoder-Decoder Models (e.g., BART)



(c) Variable Encoder-Decoder (VECO)

Comparisons	Encoder-Only	Encoder-Decoder	Variable Encoder-Decoder (VECO)
How to build representations	Jointly encode the input (x, y), i.e., let them interact with each other starting from the embedding layer	Firstly extract the "global" representation of x from the last layer of encoder, and then interact with y via extra Cross-attention	
What's the prediction task	Recover masked tokens	Generate next words	Recover masked tokens to *keep a more consistent optimization direction on shared parameters*
When applied to NLU (Stars are result ratings)	Use the whole encoder (Self-attention & FFN) ★★★★	• Only extract encoder (e.g., MMTE): ★★ • Use both encoder and decoder (e.g., BART): ★★★	Extract Self-attention & FFN to initialize an encoder: ★★★★☆
When applied to NLG (Stars are result ratings)	Can not fully initialize the decoder (e.g., BERT, XLM-R), or change the traditional generation way (e.g., UNILM): ★★★	Use the whole encoder-decoder ★★★★	Can provide complete initialization for an encoder-decoder ★★★★☆
Features	• Compatible with various NLU tasks • Compared to mainstream NLG framework, it lacks a "global" understanding of x when interacting with y	• Compatible with mainstream NLG framework • Can not beat the encoder-only models on NLU tasks, even with more computation and memory	• *Flexibility*: Easily applicable to both encoder and encoder-decoder downstream frameworks with the most streamlined parameters • *Mutuality*: Enables NLU and NLG to boost each other, and achieve the SOTA results on various NLU and NLG tasks