

\_a \_cute \_cat \_is \_pla ying \_with \_a \_toy \_b anana

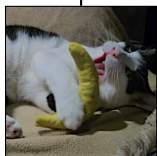
Autoregressive Transformer Language Model

Input  
Vectors

Linear Projection

Transformer Vision Encoder

Image  
Patches



Image

Linear Projection (Embedding)

\_a \_cute \_cat \_is \_pla

Subword Tokenizer

"a cute cat is playing with a toy banana"