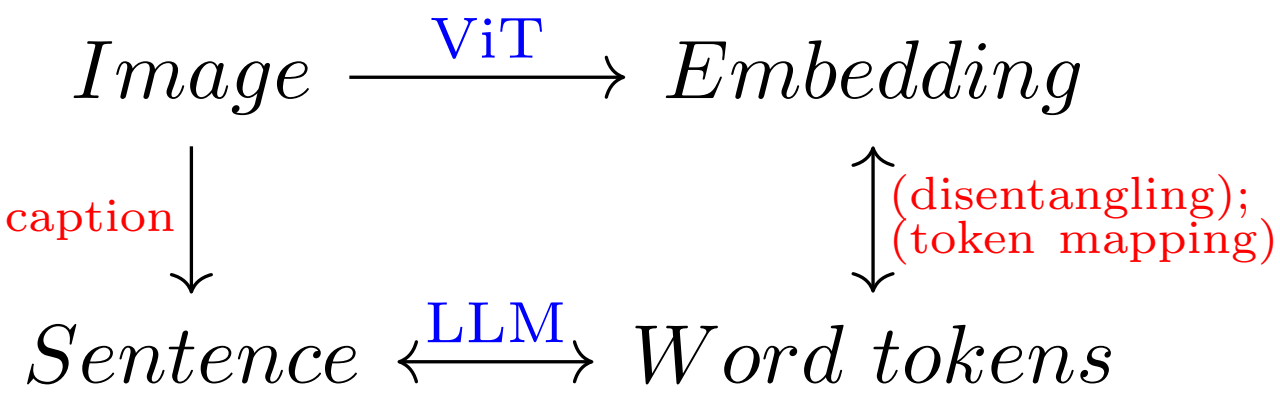
Image
ViT
Embedding
caption
(disentangling);
(token mapping)
Sentence
LLM
Word tokens