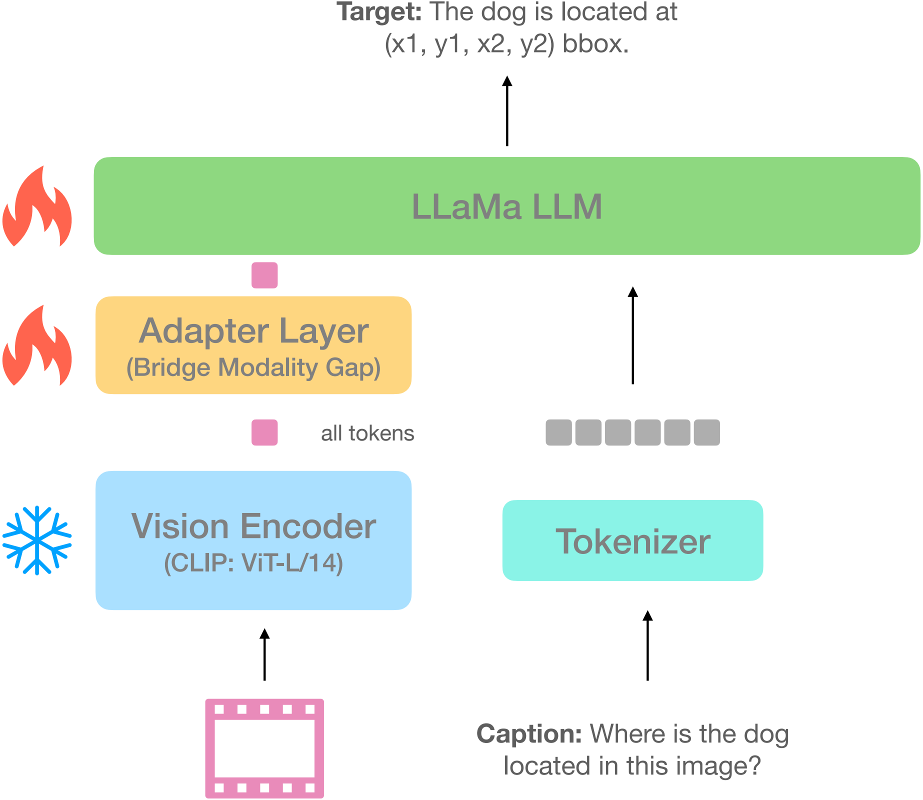
Target: The dog is located at
(x1, y1, x2, y2) bbox.
LLaMa LLM
Adapter Layer
(Bridge Modality Gap)
all tokens
Vision Encoder
(CLIP: ViT-L/14)
Tokenizer
Caption: Where is the dog
located in this image?