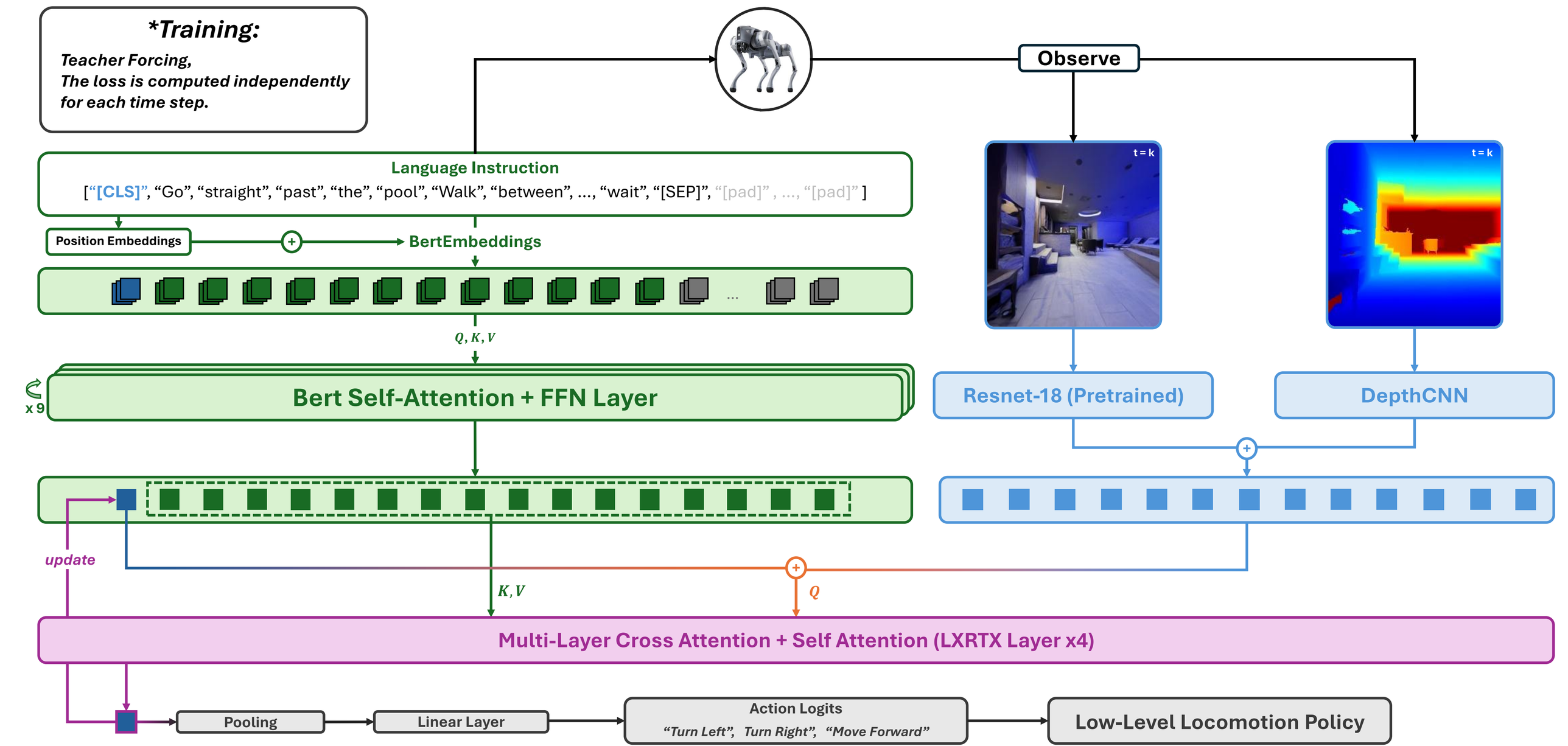
*Training:
Teacher Forcing,
The loss is computed independently
for each time step.
Observe
t = k
t = k
Language Instruction
["[CLS]", "Go", "straight", "past", "the", "pool", "Walk", "between", ..., "wait", "[SEP]", "[pad]" , ..., "[pad]"]
Position Embeddings
BertEmbeddings
Q, K, V
x 9
Bert Self-Attention + FFN Layer
Resnet-18 (Pretrained)
DepthCNN
update
K, V
Q
Multi-Layer Cross Attention + Self Attention (LXRTX Layer x4)
Pooling
Linear Layer
Action Logits
"Turn Left", "Turn Right", "Move Forward"
Low-Level Locomotion Policy