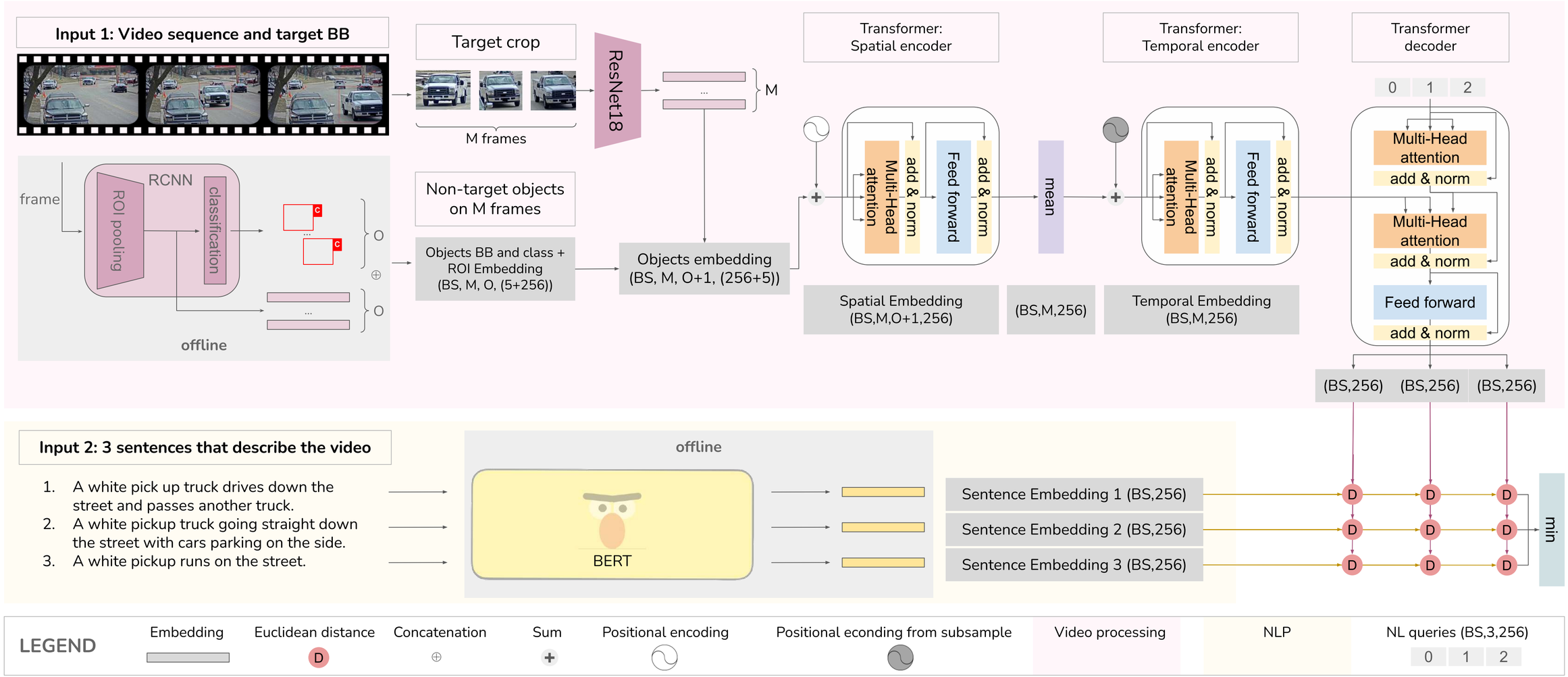
Input 1: Video sequence and target BB
Target crop
M frames
ResNet18
M
frame
RCNN
ROI pooling
classification
O
O
offline
Non-target objects on M frames
Objects BB and class + ROI Embedding (BS, M, O, (5+256))
Objects embedding (BS, M, O+1, (256+5))
Transformer: Spatial encoder
Multi-Head attention
add & norm
Feed forward
add & norm
Spatial Embedding (BS,M,O+1,256)
mean
(BS,M,256)
Transformer: Temporal encoder
Multi-Head attention
add & norm
Feed forward
add & norm
Temporal Embedding (BS,M,256)
Transformer decoder
0 1 2
Multi-Head attention
add & norm
Multi-Head attention
add & norm
Feed forward
add & norm
(BS,256) (BS,256) (BS,256)
Input 2: 3 sentences that describe the video
1. A white pick up truck drives down the street and passes another truck.
2. A white pickup truck going straight down the street with cars parking on the side.
3. A white pickup runs on the street.
offline
BERT
Sentence Embedding 1 (BS,256)
Sentence Embedding 2 (BS,256)
Sentence Embedding 3 (BS,256)
D
min
LEGEND
Embedding
Euclidean distance
Concatenation
Sum
Positional encoding
Positional econding from subsample
Video processing
NLP
NL queries (BS,3,256)
0 1 2