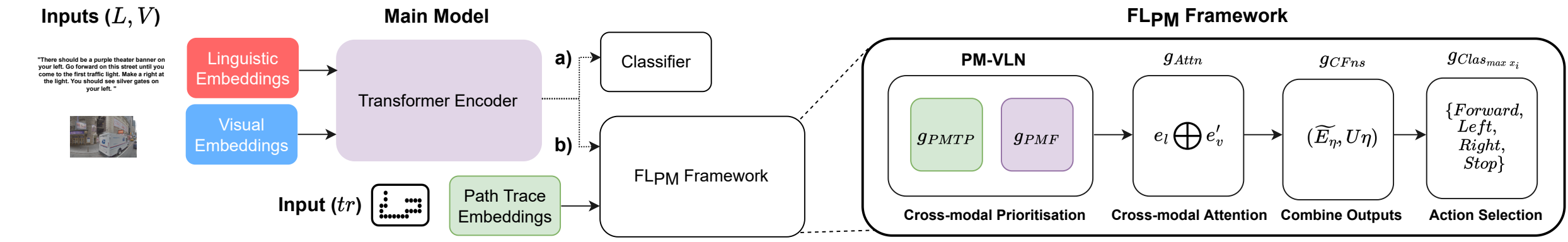
Inputs ($L, V$)
"There should be a purple theater banner on your left. Go forward on this street until you come to the first traffic light. Make a right at the light. You should see silver gates on your left. "
Linguistic Embeddings
Visual Embeddings
Main Model
Transformer Encoder
a)
Classifier
b)
FLPM Framework
Input ($tr$)
Path Trace Embeddings
FLPM Framework
PM-VLN
$g_{PMTP}$
$g_{PMF}$
Cross-modal Prioritisation
$g_{Attn}$
$e_l \bigoplus e'_v$
Cross-modal Attention
$g_{CFns}$
$(\widetilde{E_\eta}, U\eta)$
Combine Outputs
$g_{Clas_{max\ x_i}}$
{Forward, Left, Right, Stop}
Action Selection