Empower Words: DualGround for Structured Phrase and Sentence-Level Temporal Grounding

Minseok Kang
Yonsei University
louis0503@yonsei.ac.kr

Minhyeok Lee
Yonsei University
hydragon516@yonsei.ac.kr

Minjung Kim
LG Electronics
Minjung12.kim@lge.com

Donghyeong Kim
Yonsei Universiy
2donghyung87@yonsei.ac.kr

Sangyoun Lee
Yonsei University
syleee@yonsei.ac.kr

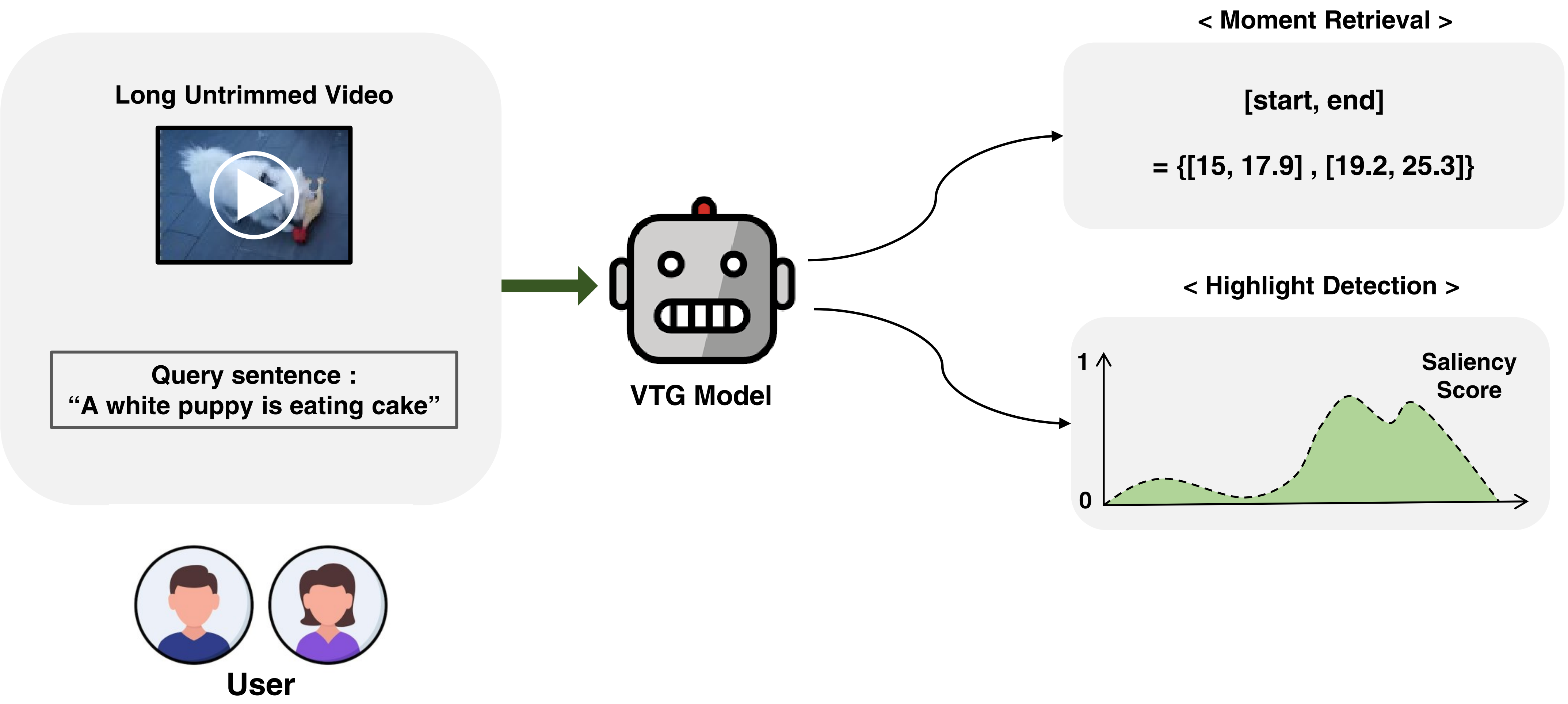
NEURAL INFORMATION PROCESSING SYSTEMS

YONSEI UNIVERSITY

MVP LAB
Yonsei Image and Video Pattern Recognition Lab

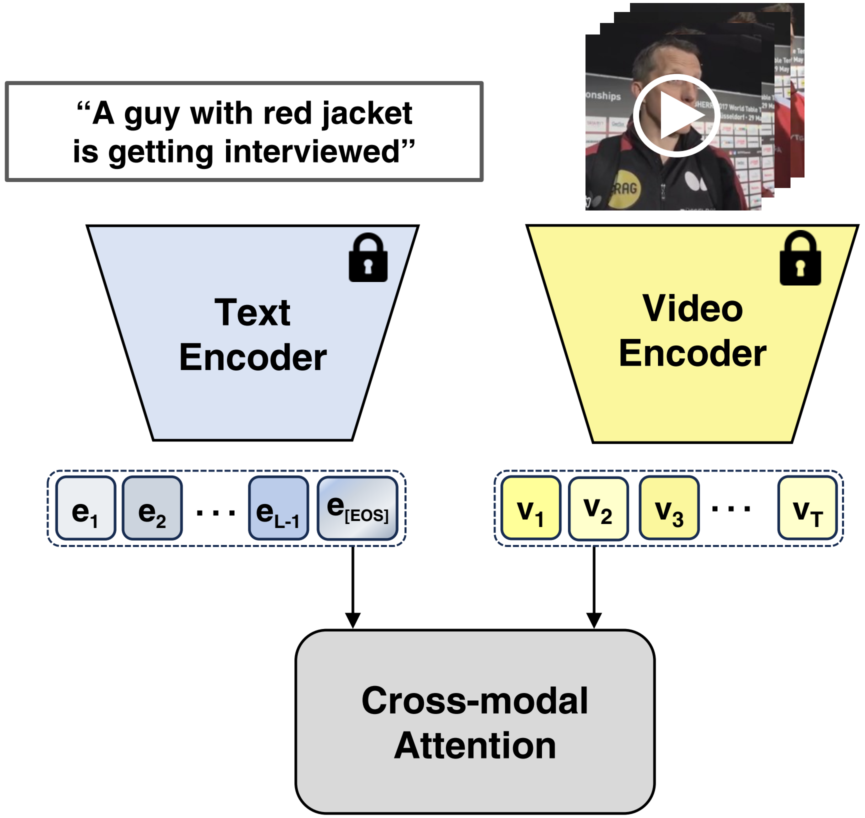
Video Temporal Grounding

"What is Video Temporal Grounding?"



Introduction

“ [EOS] over-reliance phenomenon”



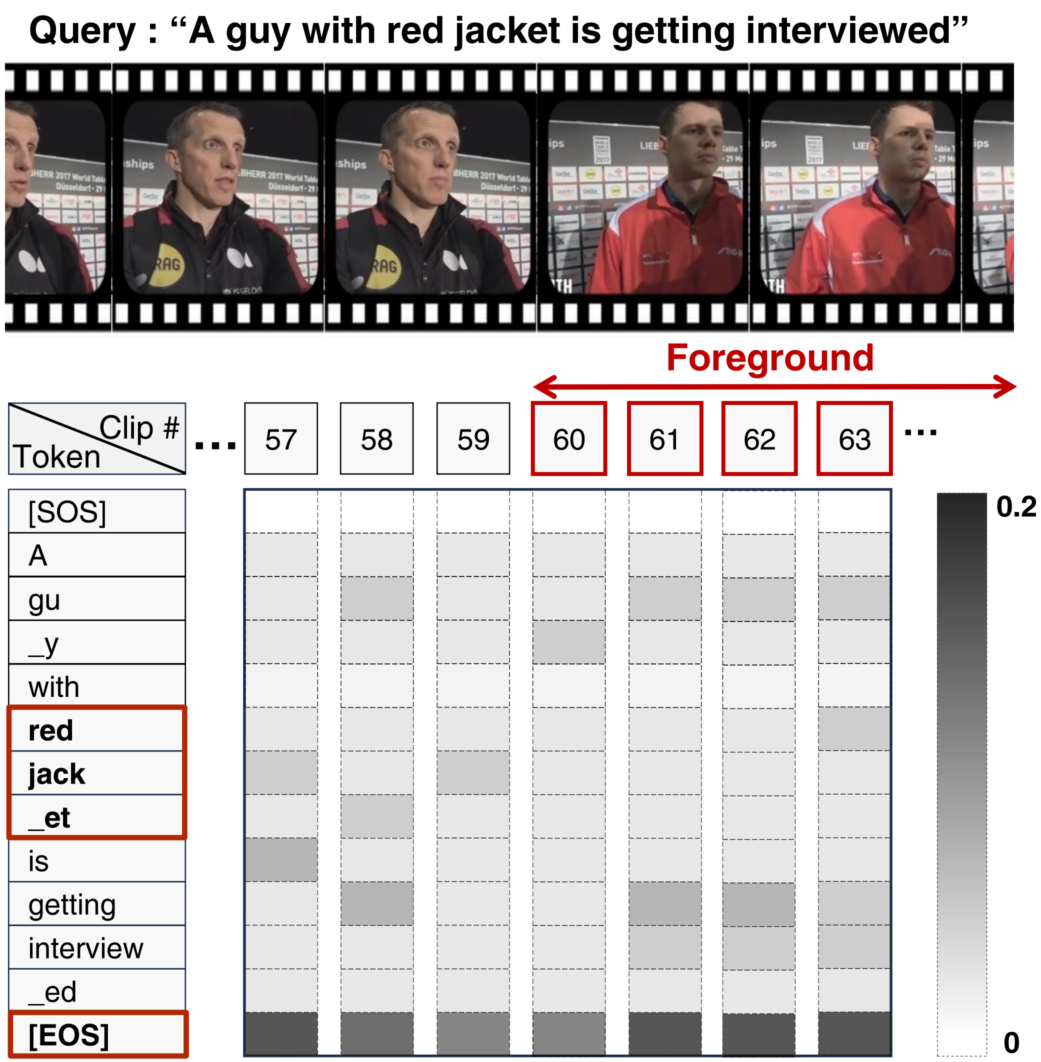
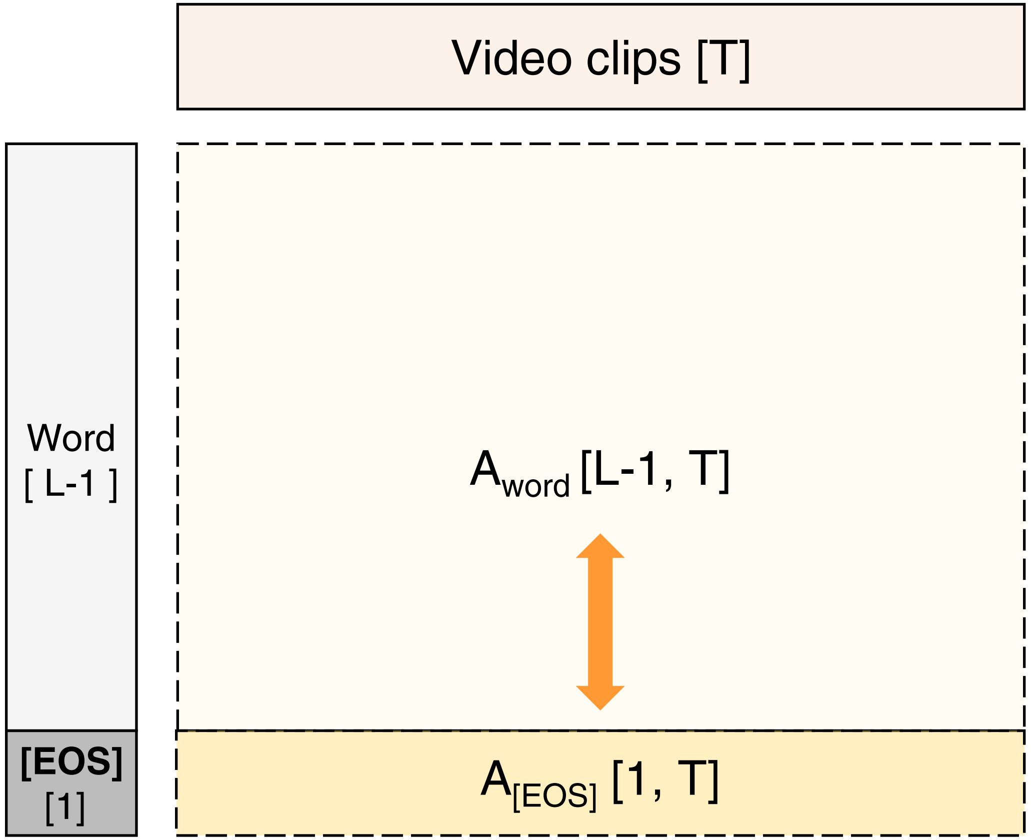


Figure 1(b). Cross-attention map from FlashVTG

Introduction

Quantitative analysis of [EOS] reliance phenomenon



Cross-modal Attention Map

Pearson Correlation

$$r_i = \frac{\Sigma(A_i - \bar{Y}_i)\left(A_{[EOS]} - \bar{Y}_{[EOS]}\right)}{\sqrt{\left(\Sigma(A_i - \bar{Y}_i)^2 \Sigma\left(A_{[EOS]} - \bar{Y}_{[EOS]}\right)^2\right)}} \qquad \bar{r} = \frac{1}{L-1} \sum_{i \neq [EOS]} r_i$$

Spearman Rank Correlation

$$\rho_i = 1 - \left(6\Sigma_t d_{i,t}^2\right) / \left(T(T^2 - 1)\right)$$

$$d_{i,t} = rank\left(A_{i,t}\right) - rank\left(A_{EOS,t}\right) \qquad \bar{\rho} = \frac{1}{L-1} \sum_{i \neq [EOS]} \rho_i$$

Model	Train		Val	
	Pearson	Spearman	Pearson	Spearman
CG-DETR	0.8960	0.8914	0.5962	0.7622
TR-DETR	0.8110	0.7754	0.6021	0.6340
FlashVTG	**0.9745**	**0.9801**	**0.6771**	**0.7800**

Table A5. Correlation between [EOS] and word-token attentions

Introduction

Higher correlation : [EOS] ≈ Word + [EOS]

Lower correlation : [EOS] >> Word + [EOS]

EOS dominant state is VTG model's current optimum

"Does [EOS] really capture everything?"

Not always — [EOS] isn't visually grounded, salient visual cues can be under-represented.

➔ Word tokens can complement [EOS].

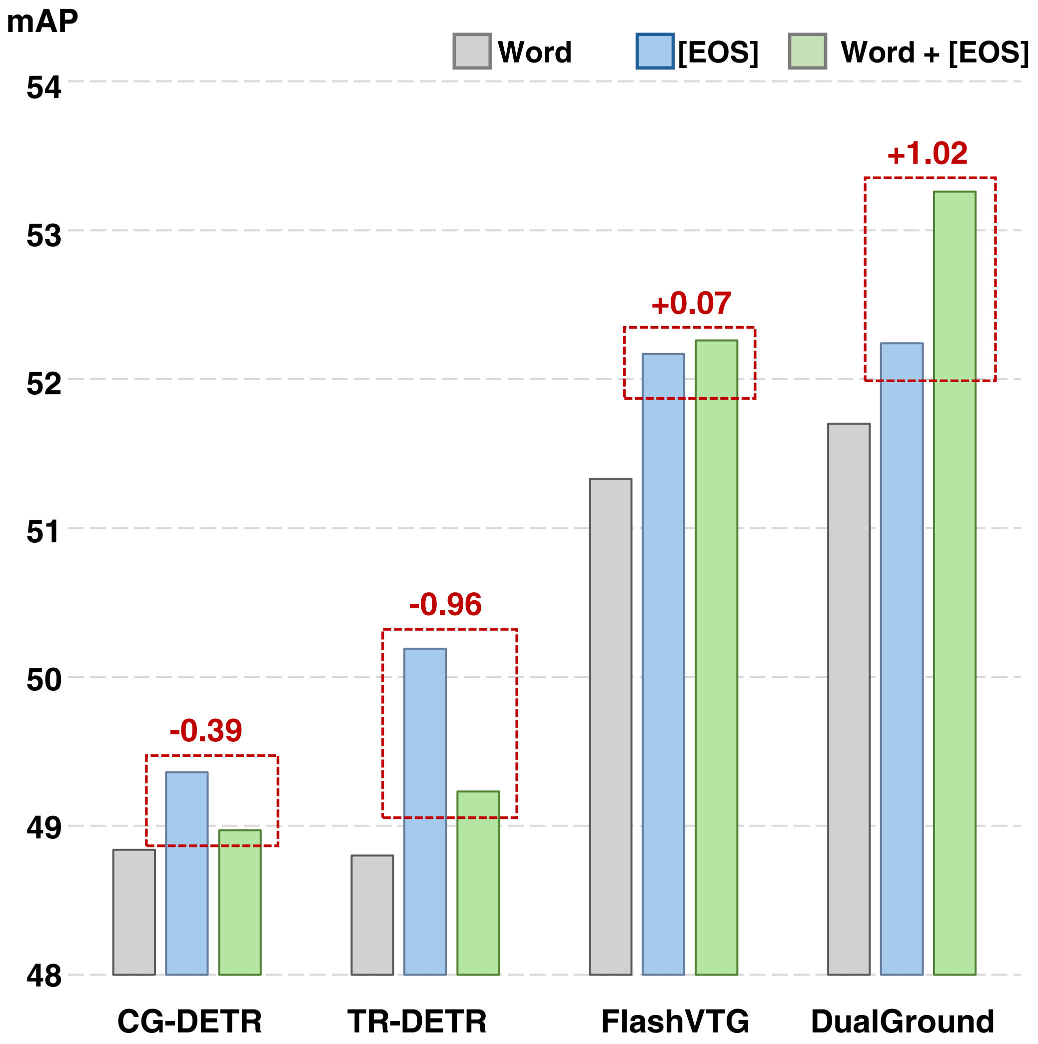


Figure 1(a). MR performance comparison with token configurations

Method

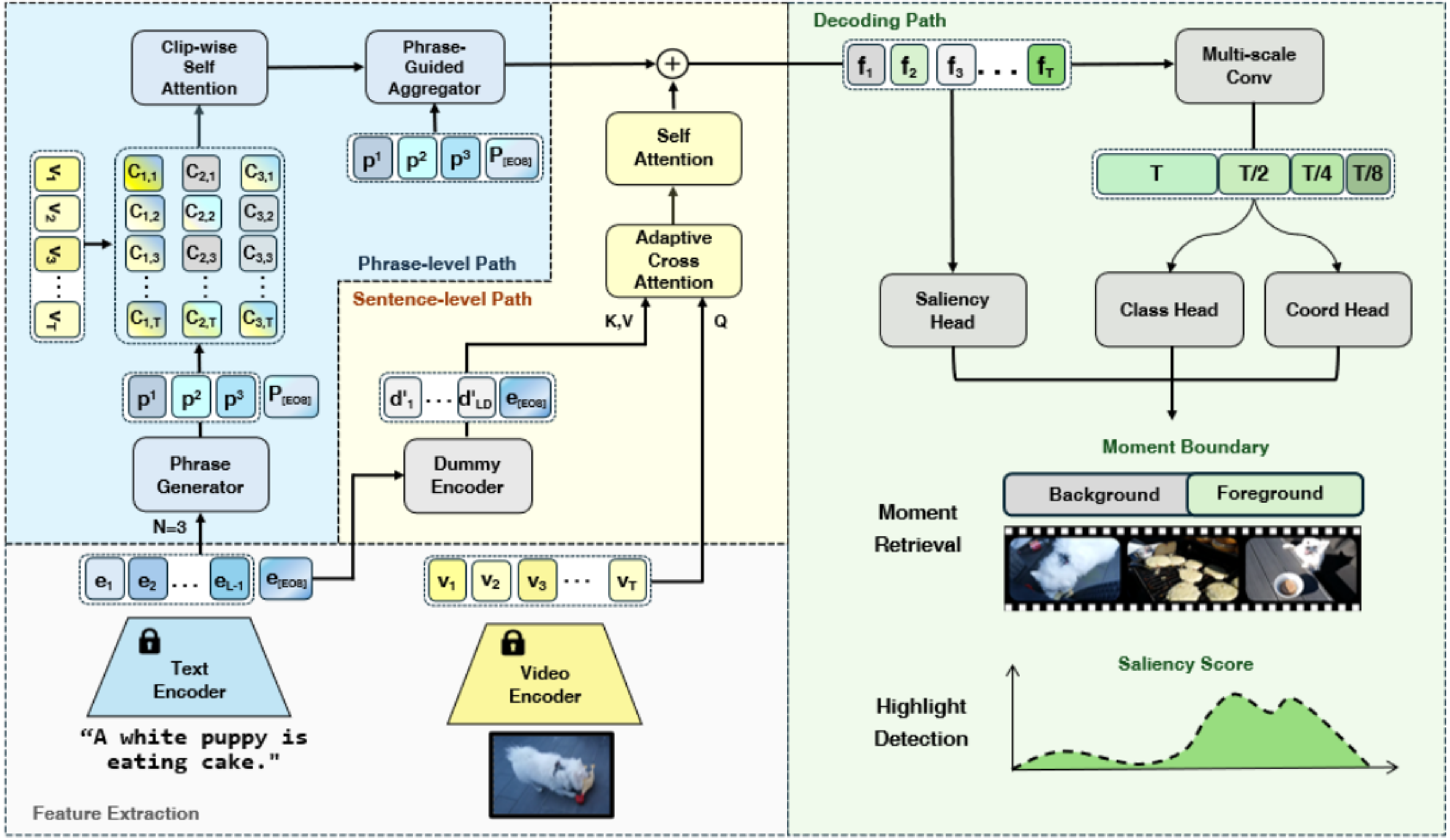
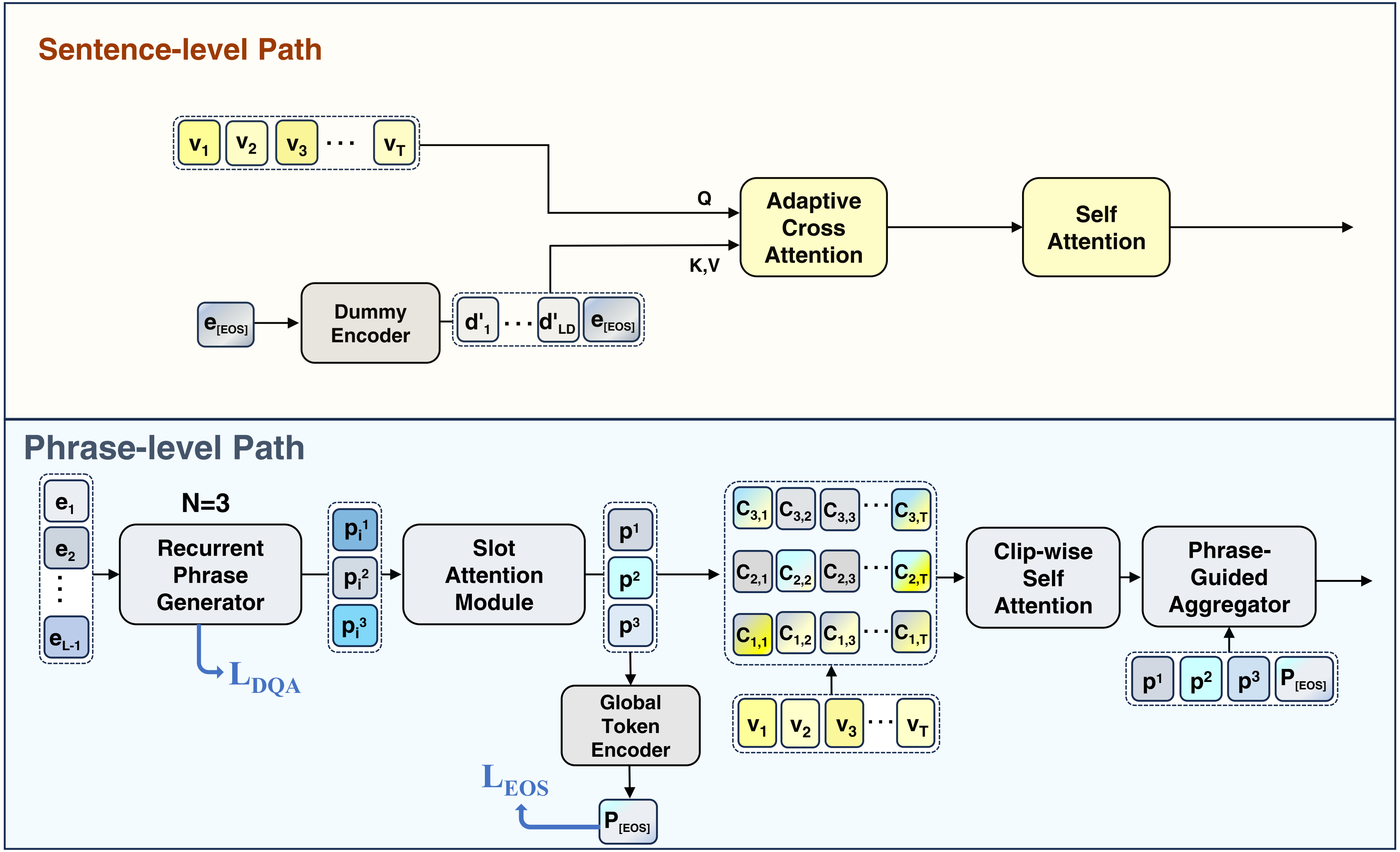


Figure 2. Overall architecture of DualGround

Method



Sentence-level Path
v1 v2 v3 ··· vT
Q
K,V
Adaptive Cross Attention
Self Attention
e[EOS]
Dummy Encoder
d'1 ··· d'LD e[EOS]
Phrase-level Path
e1
e2
eL-1
N=3
Recurrent Phrase Generator
LDQA
pi1
pi2
pi3
Slot Attention Module
p1
p2
p3
Global Token Encoder
LEOS
P[EOS]
C3,1 C3,2 C3,3 ··· C3,T
C2,1 C2,2 C2,3 ··· C2,T
C1,1 C1,2 C1,3 ··· C1,T
v1 v2 v3 ··· vT
Clip-wise Self Attention
Phrase-Guided Aggregator
p1 p2 p3 P[EOS]

Experiment Result

Table1. MR & HD results on QVHihglights Test split

Method	Backbone	R1@0.5	R1@0.7	mAP	mAP@0.5	mAP@0.75	VG-mAP	VG-Hit@1
CG-DETR [18]	SF+C	65.43	48.38	42.86	64.51	42.77	40.33	66.21
TR-DETR [24]	SF+C	64.66	48.96	42.62	63.98	43.73	39.91	63.42
UVCOM [31]	SF+C	63.55	47.47	43.18	63.37	42.67	39.74	64.20
R2-Tuning [13]	C	68.03	49.35	46.17	69.04	47.56	40.75	64.20
FlashVTG [1]	SF+C	66.08	50.00	48.70	67.99	47.59	41.07	66.10
DualGround	SF+C	**68.20**	**51.72**	**49.02**	**69.23**	**47.71**	**41.15**	**66.30**
FlashVTG [1]	IV2	70.69	53.96	52.00	72.33	53.85	**44.09**	**71.00**
DualGround	IV2	**71.87**	**56.94**	**52.73**	**72.41**	**54.38**	44.02	70.80

Table2. MR & HD results on QVHihglights Val split

Method	Backbone	R1@0.5	R1@0.7	mAP	mAP@0.5	mAP@0.75	VG-mAP	VG-Hit@1
CG-DETR [18]	IV2	70.06	55.87	48.93	69.85	49.56	42.30	68.71
TR-DETR [24]	IV2	71.72	55.93	48.93	70.87	50.14	43.74	70.84
FlashVTG [1]	IV2	71.48	56.06	52.61	72.37	55.03	44.08	71.48
DualGround	IV2	**73.48**	**58.97**	**53.26**	**72.99**	**56.35**	**44.12**	**71.62**

Table5. MR results on Charades Test split

Method	Backbone	R1@0.5	R1@0.7
UniVTG [11]	C+SF	59.25	36.64
CG-DETR [18]	C+SF	58.41	36.32
TR-DETR [24]	C+SF	57.61	33.52
FlashVTG [1]	C+SF	61.08	37.89
DualGround	C+SF	**61.11**	**38.52**
CG-DETR [18]	IV2	70.40	48.40
TR-DETR [24]	IV2	69.73	46.33
FlashVTG [1]	IV2	70.32	49.87
DualGround	IV2	**70.67**	**50.33**

Experiment Result

Phrase-Word Attention map

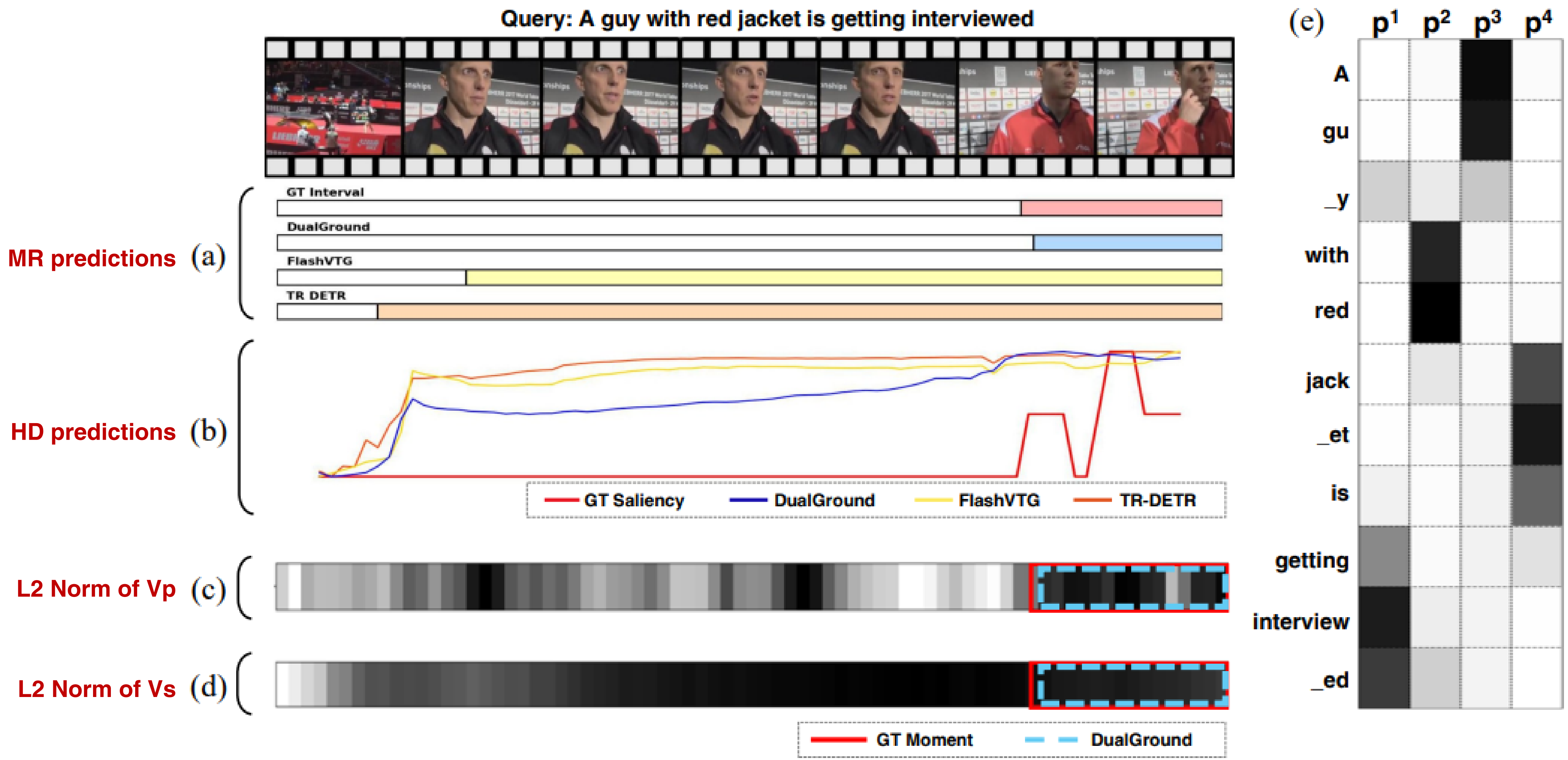


Figure 3. Visualization results on QVHighlights val split

Thank You