Query-Dependent Video Representation for Moment Retrieval and Highlight Detection

WonJun Moon[1], Sangeek Hyun[1], SangUk Park[2], Dongchan Park[2], Jae–Pil Heo[1]

Sungkyunkwan University[1], Pyler[2]

Paper ID : 6761
THU-PM-231

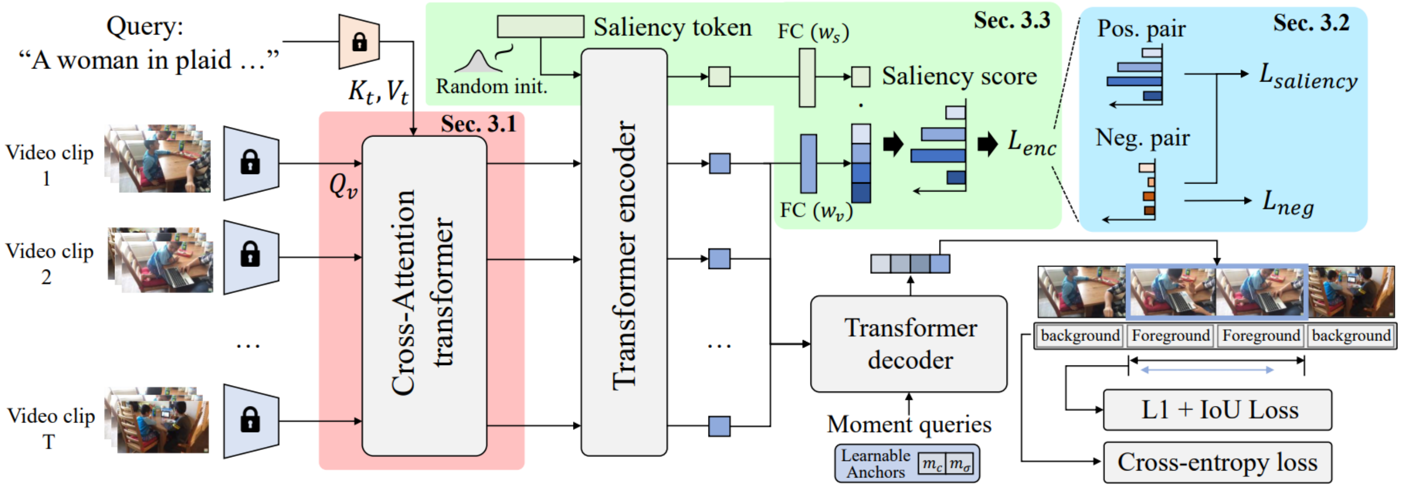
Abstract

Query-Dependent DETR for Video Moment Retrieval & Highlight Detection

Given a **Football Video**

A person wants to see the moment
"A man scored the winning goal."

- In previous works, video and text query are forwarded to self-attention layers without the consideration of importance of similarity between per-frame and query.



Query-Dependent Video Representation
Learning negative pair to reduce modality gap
Input-Adaptive Saliency Predictor

Introduction

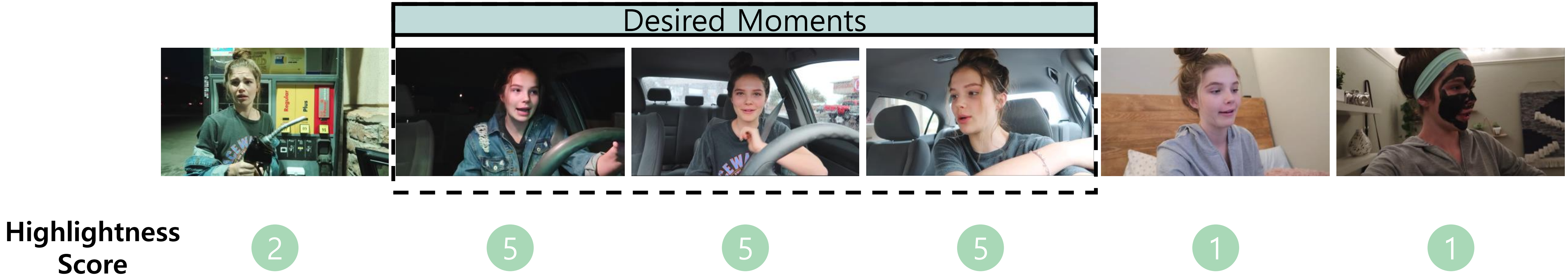
Video Moment Retrieval & Highlight Detection

- Video sources are often very long that it is hard to capture the desired moments.
- We need an automatic tool to assist finding the desired moments.

Introduction

Moment Retrieval

Given the text description for desired moments : "***A girl speaking from her car*** ",
Moment Retrieval is to clip the desired moments.

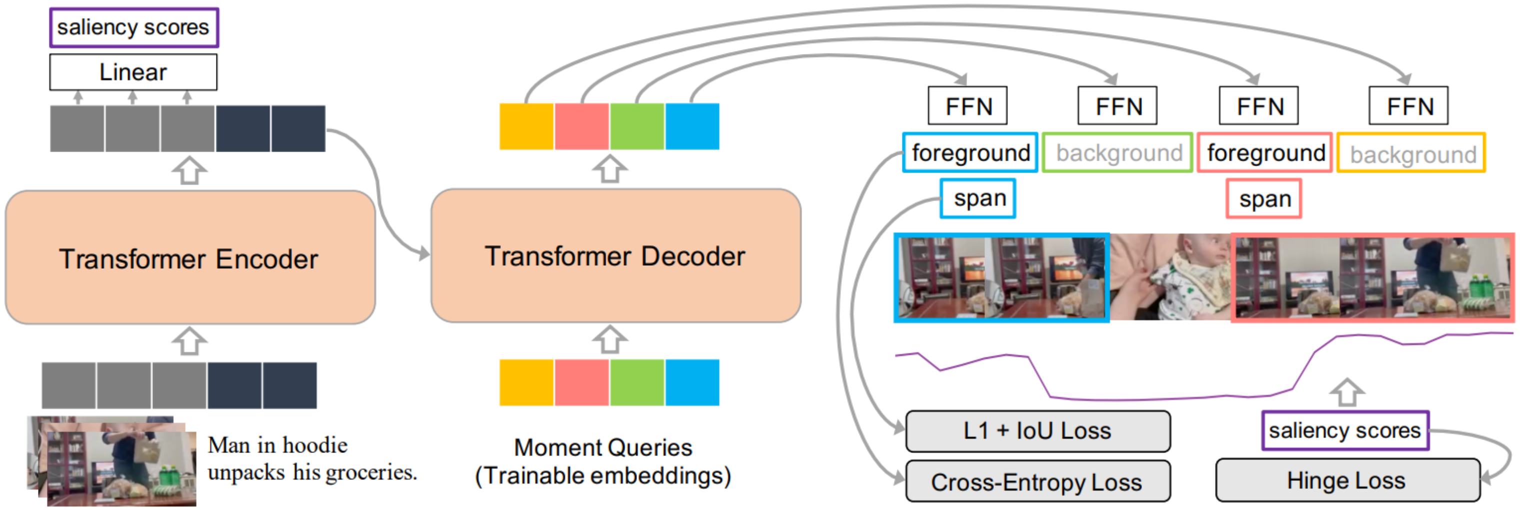


Highlight Detection

Supervised with the human annotated highlightness scores,
Highlight Detection is to learn the **frame-wise highlightness** in the human perspective.

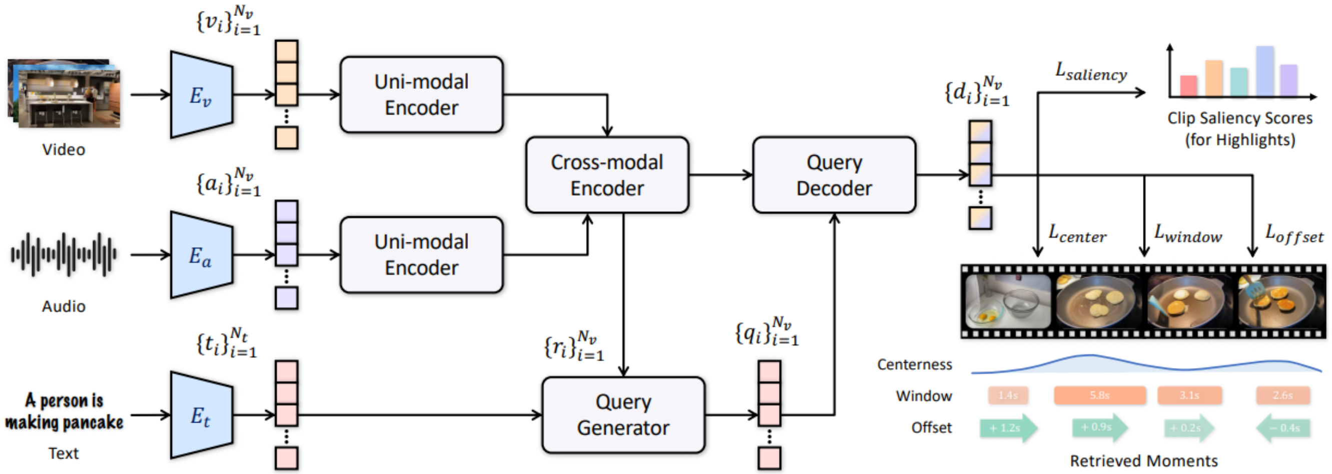
Background

Previous works



Moment- Detr (NeurIPS 2021)

UMT (CVPR 2022)

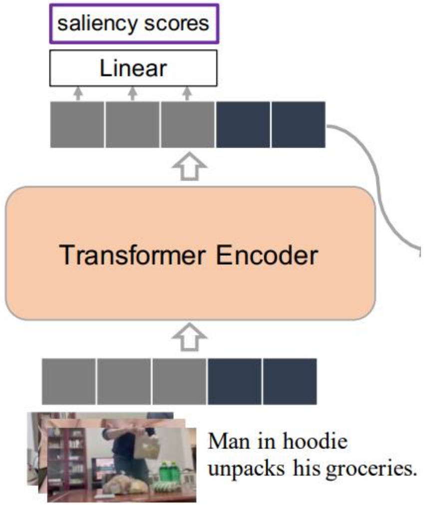


Background

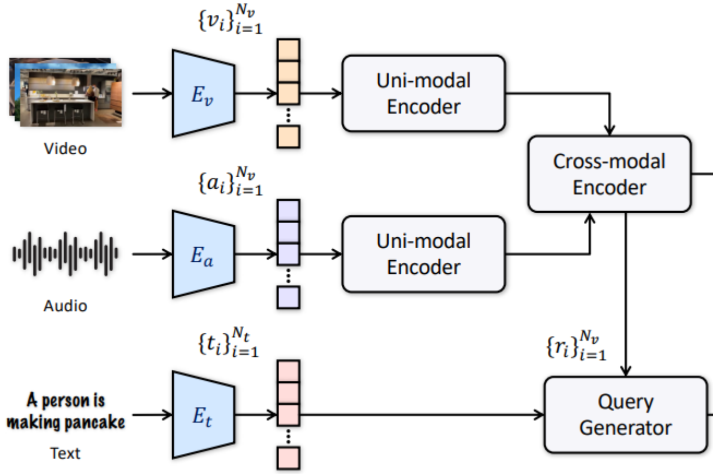
Motivation

- **Text query, the description for desired moments, are overlooked while extracting video representation.**

Moment- Detr (NeurIPS 2021)



UMT (CVPR 2022)

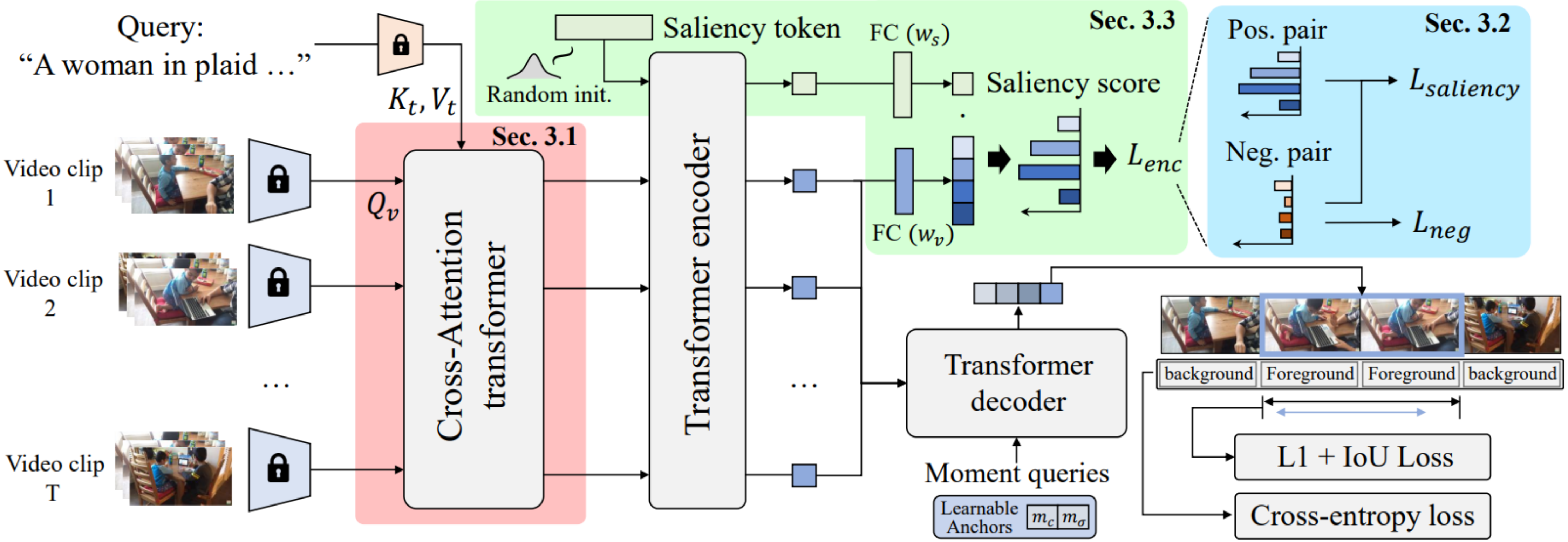


Due to modality gap, video features are more likely to be utilized in attention layers than textual information.

Due to self-attention modules before query insertion, each frame feature may no longer depict frame-specific information but video-descriptive features.

Overview

Query-Dependent DETR



Query-Dependent Video Representation

Input-Adaptive Saliency Predictor

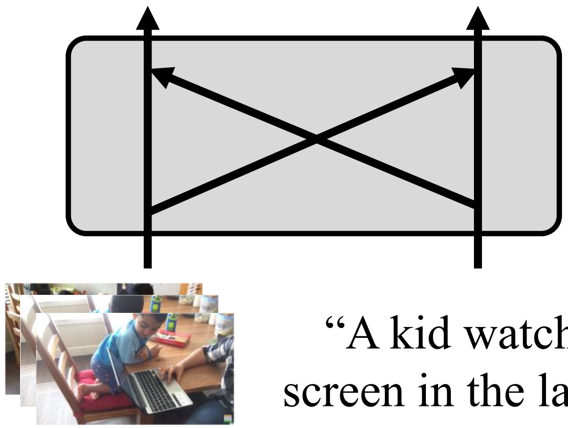
Reducing Modality GAP

Approach – Cross-Attention Transformer Encoder

Cross-Attention Transformer Encoder (CATE)

- Previous works (e.g. Moment-DETR) struggle to learn to relation between video and text query
- Adopting cross-attention on very first layer of transformer encoder
- Cross-Attention Transformer Encoder ensures the consistency contribution of text on video representation

Self-Attention

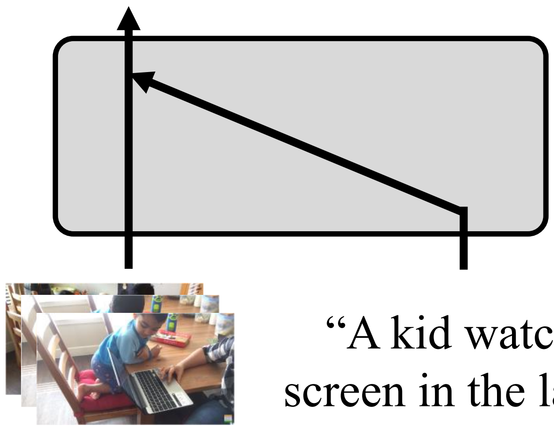


"A kid watch the screen in the laptop."

Video input **Text query**

- Interaction between video and text rep.
- Text condition may not be ensured on every clip

Cross-Attention



"A kid watch the screen in the laptop."

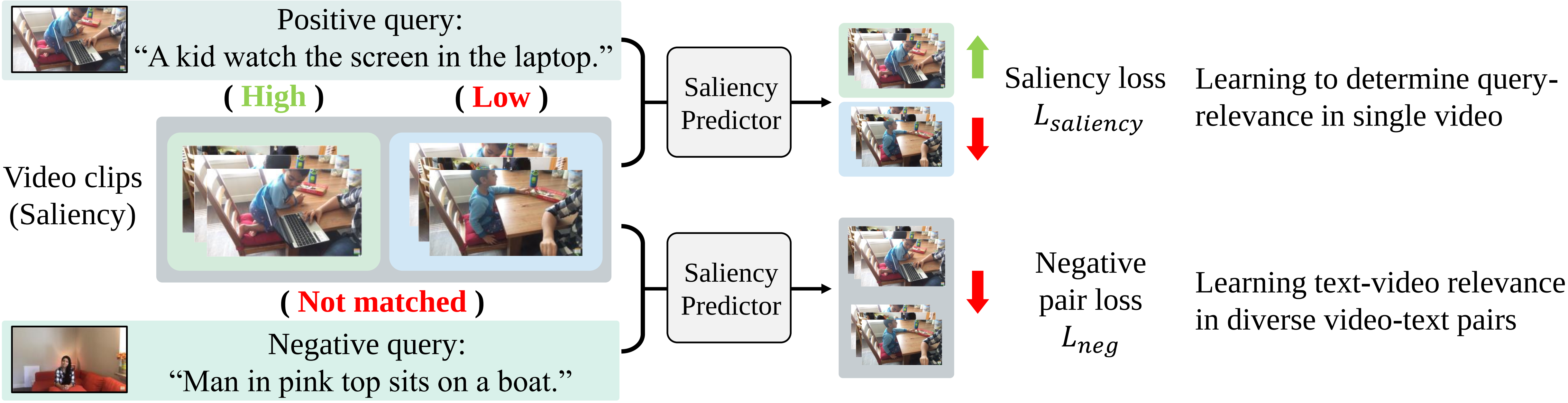
Video input **Text query**

- Simplex Interaction between video and text rep.
- Consistent text condition is ensured on every clip

Approach – Learning from Negative Relationship

Introducing Negative-relation Learning

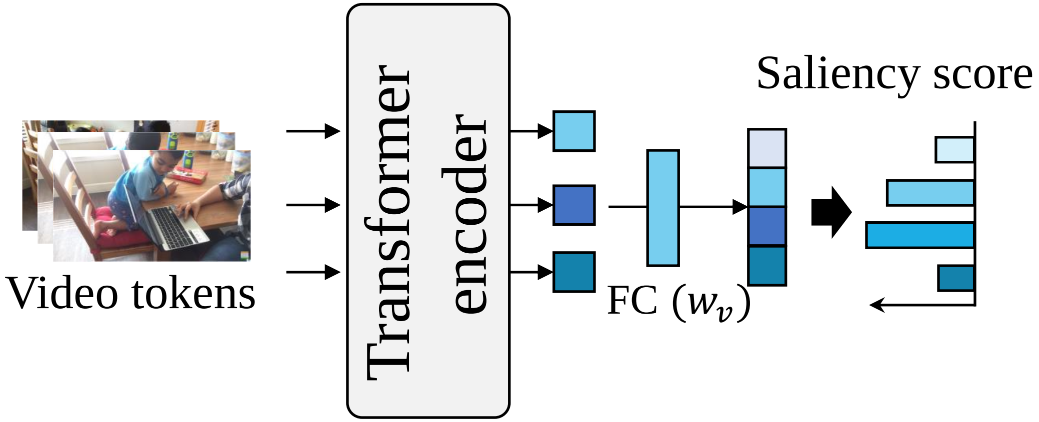
- Previous highlight-detection focus on learning the video-query relationship only with matched pairs
- Since video frames share similar appearances and similarities to a query will not be highly distinguishable, the involvement of textual information may not be high
- By penalizing the irrelevant (negative) video-query pairs, the model is encouraged to learn the general relationship between video and text queries



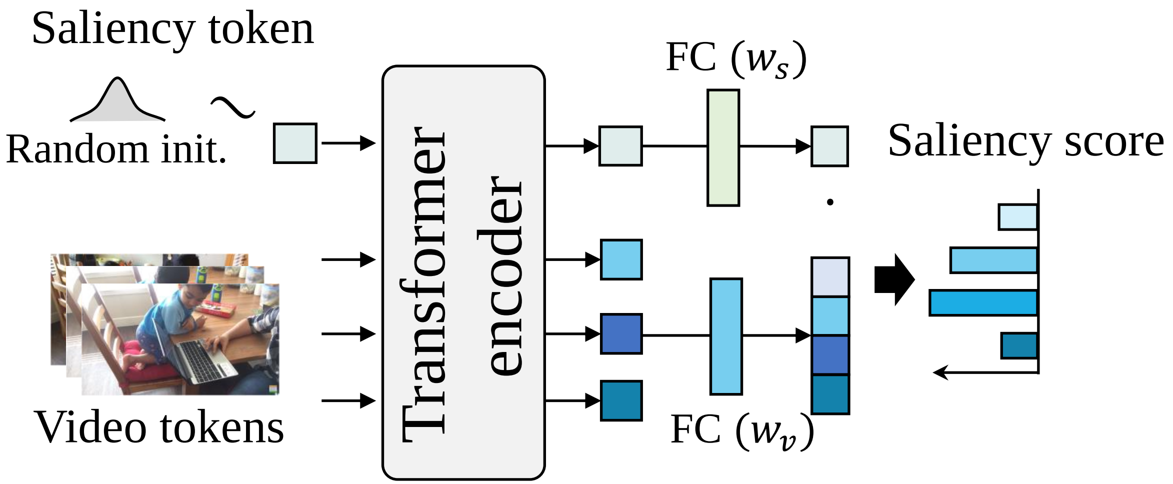
Approach – Input-Adaptive Saliency Predictor

Input-Adaptive Saliency Predictor

- Typical saliency predictor estimates the saliency score based-on single (or multiple) FC layers
- This identical criteria for the saliency prediction of every video-query pair neglects diverse nature of video-text pairs
- Introducing input-adaptive saliency predictor, which determine the saliency criteria depending on input video-text pair



Typical Saliency predictor



Input-Adaptive Saliency predictor

Experimental Results – Quantitative Results

Experiment on QVHighlights dataset (Moment Retrieval & Highlight Detection)

Method	Src	MR					HD	
		R1		mAP			>= Very Good	
		@0.5	@0.7	@0.5	@0.75	Avg.	mAP	HIT@1
BeautyThumb [47]	V	-	-	-	-	-	14.36	20.88
DVSE [35]	V	-	-	-	-	-	18.75	21.79
MCN [1]	V	11.41	2.72	24.94	8.22	10.67	-	-
CAL [13]	V	25.49	11.54	23.40	7.65	9.89	-	-
XML [29]	V	41.83	30.35	44.63	31.73	32.14	34.49	55.25
XML+ [29]	V	46.69	33.46	47.89	34.67	34.90	35.38	55.06
Moment-DETR [28]	V	$52.89_{\pm 2.3}$	$33.02_{\pm 1.7}$	$54.82_{\pm 1.7}$	$29.40_{\pm 1.7}$	$30.73_{\pm 1.4}$	$35.69_{\pm 0.5}$	$55.60_{\pm 1.6}$
QD-DETR (**Ours**)	V	$\mathbf{62.40}_{\pm 1.1}$	$\mathbf{44.98}_{\pm 0.8}$	$\mathbf{62.52}_{\pm 0.6}$	$\mathbf{39.88}_{\pm 0.7}$	$\mathbf{39.86}_{\pm 0.6}$	$\mathbf{38.94}_{\pm 0.4}$	$\mathbf{62.40}_{\pm 1.4}$
UMT [36]	V+A	56.23	41.18	53.38	37.01	36.12	38.18	59.99
QD-DETR (**Ours**)	V+A	$\mathbf{63.06}_{\pm 1.0}$	$\mathbf{45.10}_{\pm 0.7}$	$\mathbf{63.04}_{\pm 0.9}$	$\mathbf{40.10}_{\pm 1.0}$	$\mathbf{40.19}_{\pm 0.6}$	$\mathbf{39.04}_{\pm 0.3}$	$\mathbf{62.87}_{\pm 0.6}$

Experimental Results – Ablation study

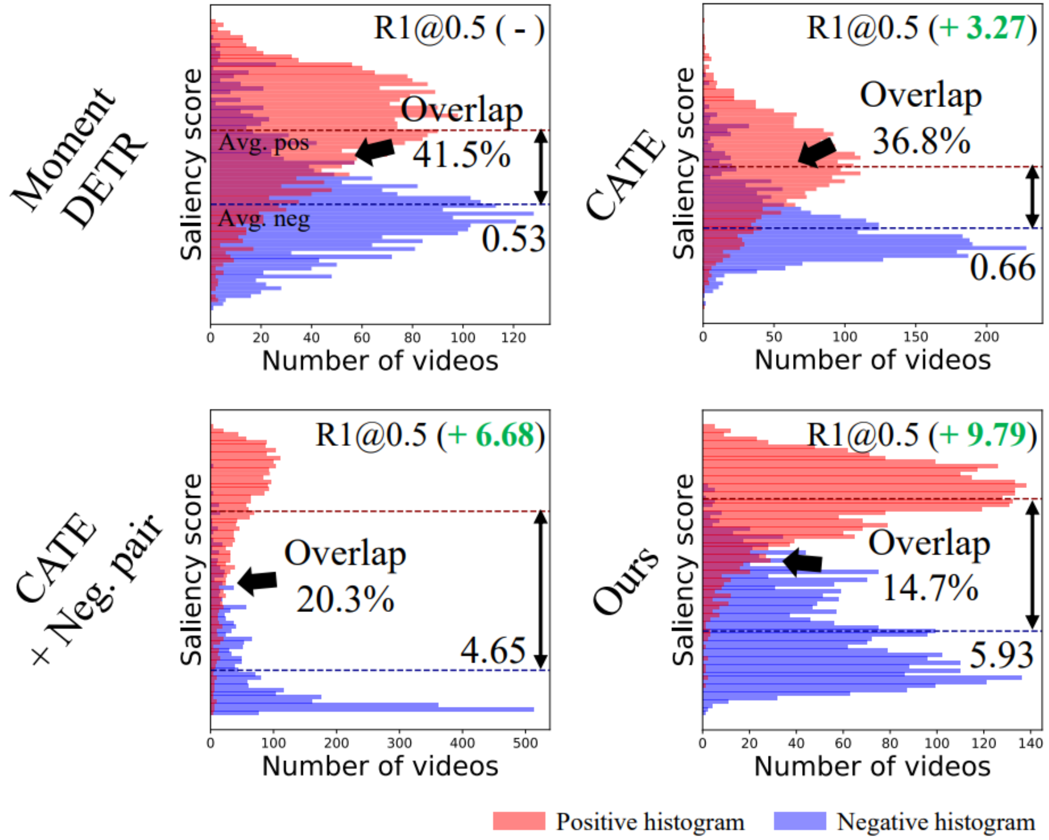
Analysis on the proposed components

Ablation study on proposed components

	CATE	Neg. Pair	ST	DAM	MR R1 @0.5	MR R1 @0.7	MR mAP Avg.	HD >= Very Good mAP	HD >= Very Good HIT@1
(a)					52.89	33.02	30.73	35.69	55.60
(b)	✓				56.16	38.71	34.07	37.14	58.34
(c)		✓			58.69	39.83	35.40	39.02	62.81
(d)			✓		55.48	37.00	32.84	37.48	58.59
(e)				✓	53.19	35.91	33.33	35.68	55.56
(f)	✓			✓	57.72	42.35	38.03	36.56	57.44
(g)	✓	✓			59.57	42.12	36.76	38.64	61.62
(h)		✓	✓		60.00	40.97	35.89	39.06	62.88
(i)	✓	✓	✓		60.32	42.39	36.93	**39.21**	62.76
(j)	✓	✓	✓	✓	**62.68**	**46.66**	**41.22**	39.13	**63.03**

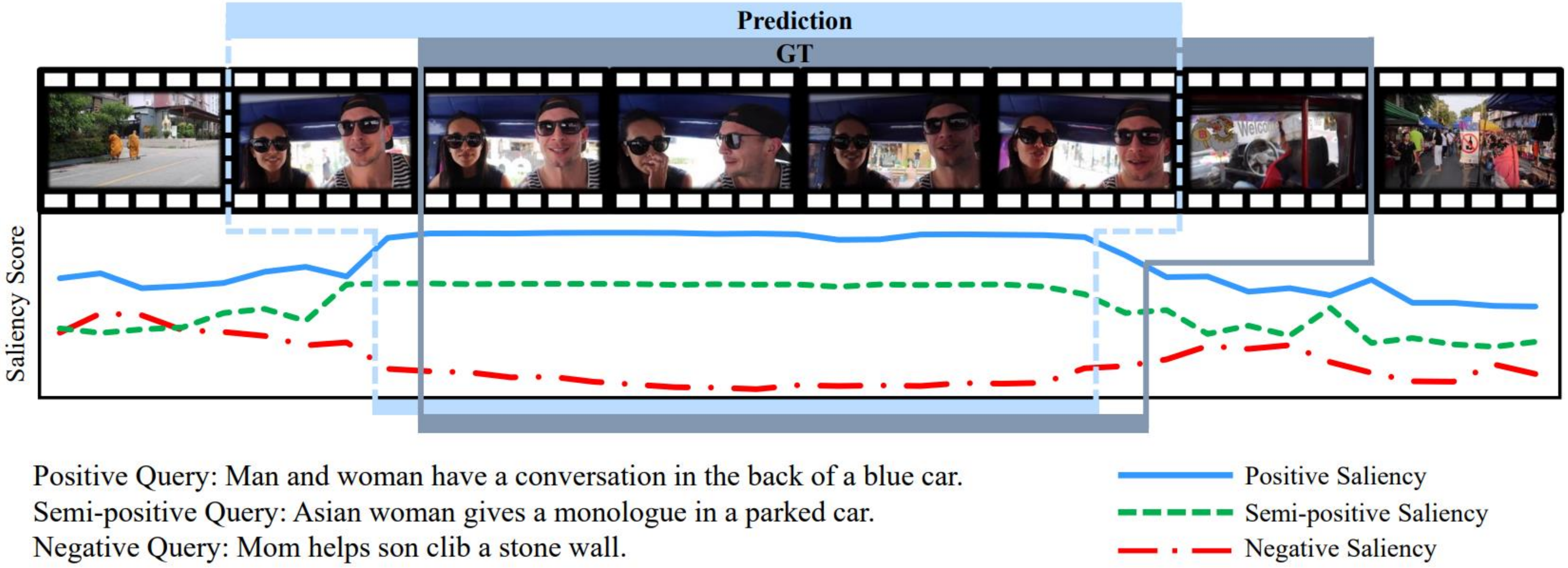
ST denotes saliency token

Saliency histogram on pos/neg pair



Experimental Results – Qualitative Result

Qualitative results – Example MR/HD prediction for given pos/semi-pos/neg pair



Contribution

- We found that the previous MR/HD methods only uses queries to play an insignificant role; they may not be capable of detecting negative queries and video-query relevance

- To tackle this issue, we introduce Query-Dependent DETR (QD-DETR) with 3 major components
    1. **Cross-Attention Transformer Encoder** to explicitly inject the context of text query into video representation.
    2. **Negative-relation learning** for encouraging the model to estimate precise accordance between video-query pairs
    3. **Input-adaptive saliency predictor** which adaptively defines the criterion of saliency scores for the given video-query pairs

- Our overall approach is verified to be superior to existing state-of-the-art methods with extensive experiments and showed that increasing the involvement of text query is essential

Thank you