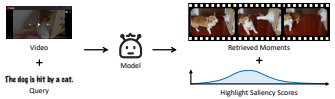
# UMT: Unified Multi-modal Transformers for Joint Video Moment Retrieval and Highlight Detection

*Ye Liu* [1] *Siyuan Li* [2] *Yang Wu* [2*] *Chang Wen Chen* [1,4] *Ying Shan* [2] *Xiaohu Qie* [3]

[1] The Hong Kong Polytechnic University [2] ARC Lab, Tencent PCG [3] Tencent PCG [4] Peng Cheng Laboratory

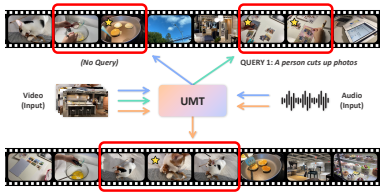
## Task Formulation

Given an untrimmed video and a natural language query, the goal of joint video moment retrieval and highlight detection is to localize all the ***moments*** in the video, in which the visual and/or audio contents are relevant to the query, while predicting ***clip-level saliency scores*** for them simultaneously.




## Motivation

Most existing works tackle the problems of video moment retrieval and highlight detection separately, ignoring the implicit relations between them. Current SOTA is also not flexible enough for multiple modality reliability situations. In this work, we propose Unified Multi-modal Transformers (UMT) to handle different modality conditions and combinations.



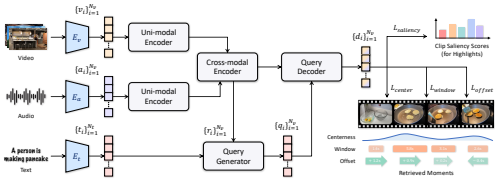

The contributions of this work are summarized as follows:

- We presented the first unified framework for joint video moment retrieval and highlight detection;
- Our model regards moment retrieval as a keypoint detection problem, largely increasing the regression accuracy;
- The effectiveness and superiority of the method have been demonstrated on diverse public datasets.

## Method Overview

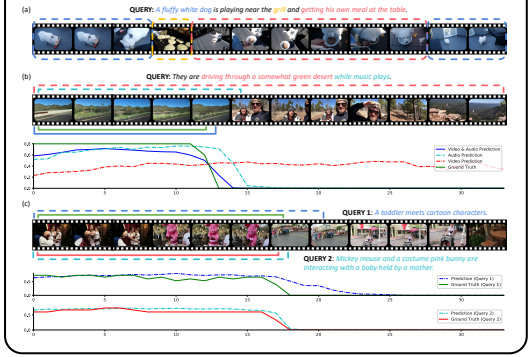
Our framework derives from the transformer encoder-decoder architecture.

- **Uni-modal Encoder:** Contextualize single-modal information;
- **Cross-modal Encoder:** Propagate multi-modal features across modalities;
- **Query Generator:** Generate clip-aligned moment queries based on text input;
- **Query Decoder:** Predict moments and highlights conditioned on the query.



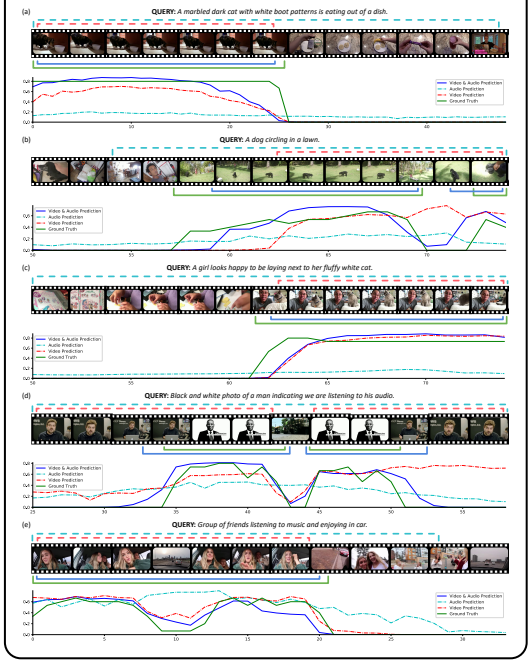

## Qualitative Analysis

The predicted moments and saliency scores are shown by ***brackets*** and ***lines***.



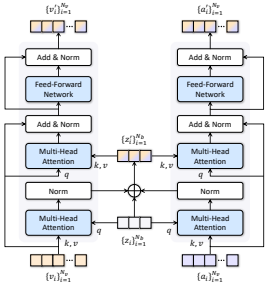

## Visualizations

The visualization results show that visual and audio features contribute to different out comes. Combining the visual and audio information can improve the performances.




## Bottleneck Transformer

We introduce bottleneck transformer to disentangle cross-modal feature aggregation into feature ***compression*** and ***expansion***, balancing performance and efficiency.




## Quantitative Analysis

| Method | MR R1 @0.5 | MR R1 @0.75 | MR mAP @0.5 | MR mAP @0.75 | MR mAP Avg. | HD ≥ Very Good mAP | HD ≥ Very Good HIT@1 |
|---|---|---|---|---|---|---|---|
| BeautyThumb | – | – | – | – | – | 14.36 | 20.88 |
| DVSE | – | – | – | – | – | 18.75 | 21.79 |
| MCN | 11.41 | 2.72 | 24.94 | 8.22 | 10.67 | – | – |
| CAL | 25.49 | 11.54 | 23.40 | 7.65 | 9.89 | – | – |
| XML | 41.83 | 30.35 | 44.63 | 31.73 | 32.14 | 34.49 | 55.25 |
| XML+ | 46.69 | 33.46 | 47.89 | 34.67 | 34.90 | 35.38 | 55.06 |
| Moment-DETR | 52.89 | 33.02 | 54.82 | 29.40 | 30.73 | 35.69 | 55.60 |
| Moment-DETR w/ PT | 59.78 | 40.33 | **60.51** | 35.36 | 36.14 | 37.43 | 60.17 |
| **UMT** (Ours) | 56.23 | 41.18 | 53.83 | 37.01 | 36.12 | 38.18 | 59.99 |
| **UMT** (Ours) w/ PT | **60.83** | **43.26** | 57.33 | **39.12** | **38.08** | **39.12** | **62.39** |

## Acknowledgements

This research is supported in part by Key-Area Research and Development Program of Guangdong Province, China with Grant 2019B010155002 and financial support from ARC Lab, Tencent PCG.