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

Ye Liu [1] Siyuan Li [2] Yang Wu [2*] Chang Wen Chen [1,4] Ying Shan [2] Xiaohu Qie [3]
[1] Department of Computing, The Hong Kong Polytechnic University
[2] Applied Research Center (ARC), Tencent PCG [3] Tencent PCG [4] Peng Cheng Laboratory
csyeliu@comp.polyu.edu.hk, changwen.chen@polyu.edu.hk
{siyuanli,dylanywu,yingsshan,tigerqie}@tencent.com

In this document, we provide more descriptions of the model architecture and implementation details to complement the main paper. Additional ablation studies and visualization on QVHighlights [2] are also incorporated to demonstrate the effectiveness of the proposed method.

1. Model Architecture

Learnable positional encodings with 0.1 dropout rates are adopted in all the encoder and decoder layers. More specifically, in uni-modal encoders, the same positional encodings are added to the Q and K matrices. In cross-modal encoders, they are added to the K matrix during feature compression and the Q matrix during feature expansion. In the query decoder, two independent positional encodings are added to the Q and K matrices, respectively.

In video- or audio-only schemes, cross-modal encoders are not necessary thus be removed. Their output normalization layers are moved to the end of the corresponding uni-modal encoders. When text queries are not available, the query generator simply outputs the visual-audio joint representations $\{r_i\}_{i=1}^{N_v}$, which will be further added with learnable positional encodings to construct moment queries.

2. Implementation Details

Similar to [1, 3], during training, each ground truth moment with quantized center $\widetilde{p}$ and window d is used to establish a 1D gaussian kernel $H_x = \exp(-\frac{(x-\widetilde{p})^2}{2{\sigma_p}^2})$ with radius r_p on the heatmap H. Here, x indicates the temporal coordinate aligned with clip indices, r_p and σ_p are window-adaptive parameters that can be computed as

$$r_p = \mu \cdot d \tag{1}$$

$$\sigma_p = \rho \cdot (r_p + 1) \tag{2}$$

where μ and ρ are hyperparameters controlling the corresponding values. We add 1 to r_p in Eq. 2 to ensure that the output σ_p is not too small, preventing extremely large values in the heatmap. We observe that the moment retrieval performance is not sensitive to these hyperparameters, thus both of them are set to 0.2 in all experiments. When testing, we compute the moment retrieval results assuming all the clips are centers to obtain a higher recall.

Following [2], we also consider the weakly-supervised pre-training on QVHighlights with automatic speech recognition (ASR) captions. During pre-training, the saliency loss is turned off since the supervision is only for moment retrieval. We observe that UMT converges much faster than Moment-DETR due to the novel formulation of moment retrieval, thus we increase the batch size to 2048 and pre-train the model for 100 epochs to prevent overfitting. Other hyperparameters strictly follow the original settings.

[*] *Corresponding author.*

Table 1. Comparison of cross-modal fusion methods on YouTube Highlights, TVSum, and QVHighlights `val` split. MR and HD denote moment retrieval and highlight detection, respectively.

Method	YouTube	TVSum	QVHighlights	
	mAP	Top-5 mAP	MR (mAP)	HD (mAP)
Concat	73.32	80.26	37.03	38.74
Mean	73.29	81.76	37.04	38.91
Sum	73.53	81.77	37.33	38.88
Bottleneck	**74.93**	**83.14**	**38.59**	**39.85**

Table 2. Ablation on number of tokens in the bottleneck transformer on QVHighlights `val` split (metric: mAP). MR and HD denote moment retrieval and highlight detection, respectively.

#Tokens	4	8	16	32	64	128
MR	**38.59**	37.19	38.29	37.15	37.20	38.18
HD	**39.85**	39.64	39.26	39.37	39.22	39.26

3. Ablation Studies

Table 1 shows the comparison among bottleneck tokens and its baselines for cross-modal feature fusion on multiple datasets. It shows that utilizing the bottleneck transformer rather than simple operations can improve the performance on both moment retrieval and highlight detection.

Table 2 studies the influence of the number of bottleneck tokens. We observe that the performance is insensitive to the number of tokens, since the feature compression and expansion process eliminate the undesirable noise.

4. Visualization

Figure 1 displays more qualitative results on QVHighlights [2]. The results show that visual and audio features contribute to different moment retrieval and highlight detection outcomes. For example, in Figure 1 (b), the video-only model fails to refine the retrieved moment given the determiner ‘indicating we are listening to his audio’, while the audio-only model can not distinguish the moment boundaries. Combining the visual and audio information can effectively improve the performances on both tasks.

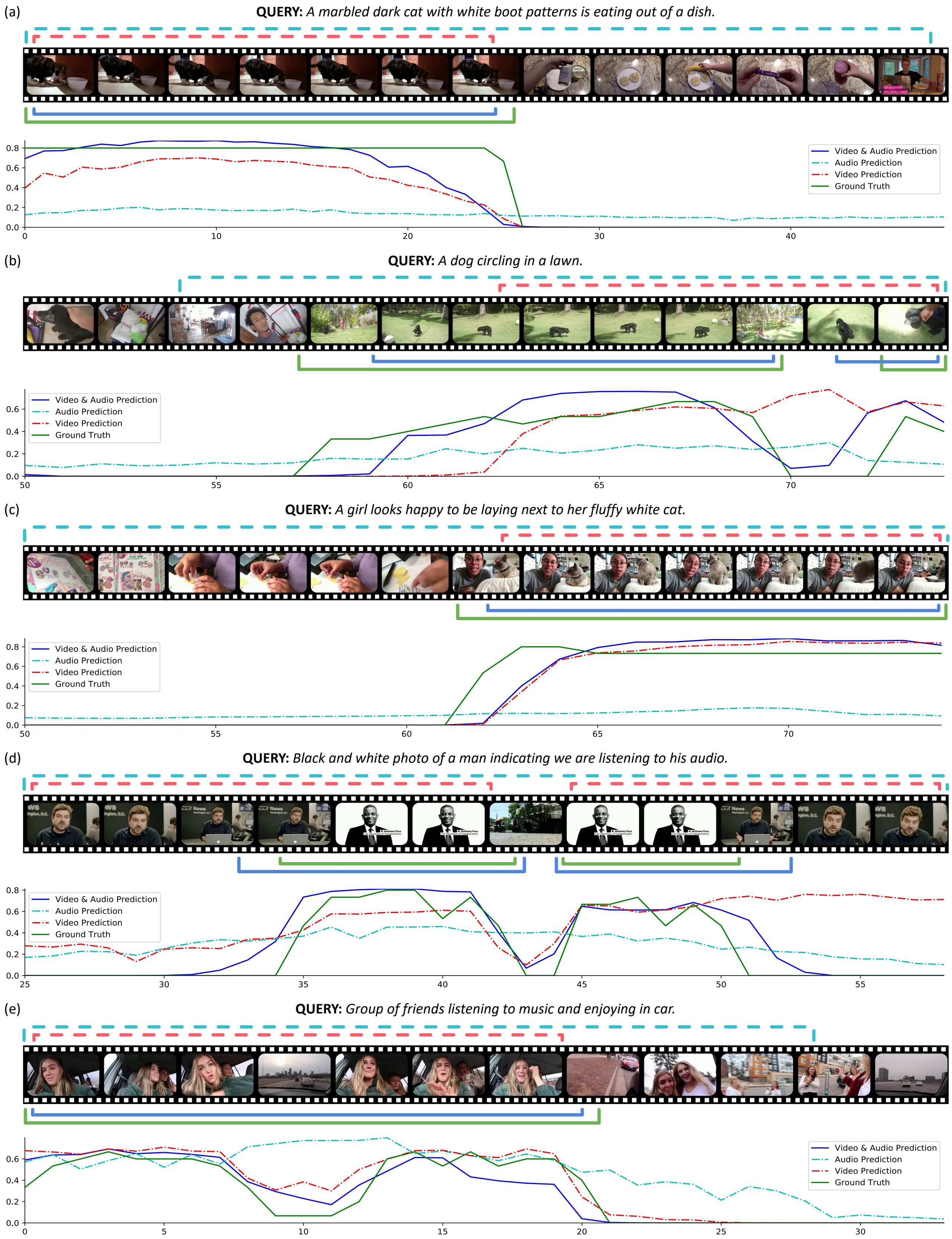
Figure 2 presents some failure cases on QVHighlights [2]. Figure 2 (a) shows that our model fails to comprehend the time adverbial clause ‘after a tiring trip’. Instead, it pays more attention to ‘a young mother and her family’ and predicts irrelevant moments. In Figure 2 (b), the visual appearances of the retrieved moment and the ground truth are similar. There are few visual clues that can be used to separate ‘shot’ and other actions. Figure 2 (c) indicates that our model understands nouns, but can not comprehend abstract words well. We argue that most failure cases are caused by the incomplete understanding of text queries, which may be remitted by using a stronger language model.

References

[1] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 734–750, 2018. 1

[2] Jie Lei, Tamara L Berg, and Mohit Bansal. Qvhighlights: Detecting moments and highlights in videos via natural language queries. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1, 2

[3] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1



(a) **QUERY:** *A marbled dark cat with white boot patterns is eating out of a dish.*

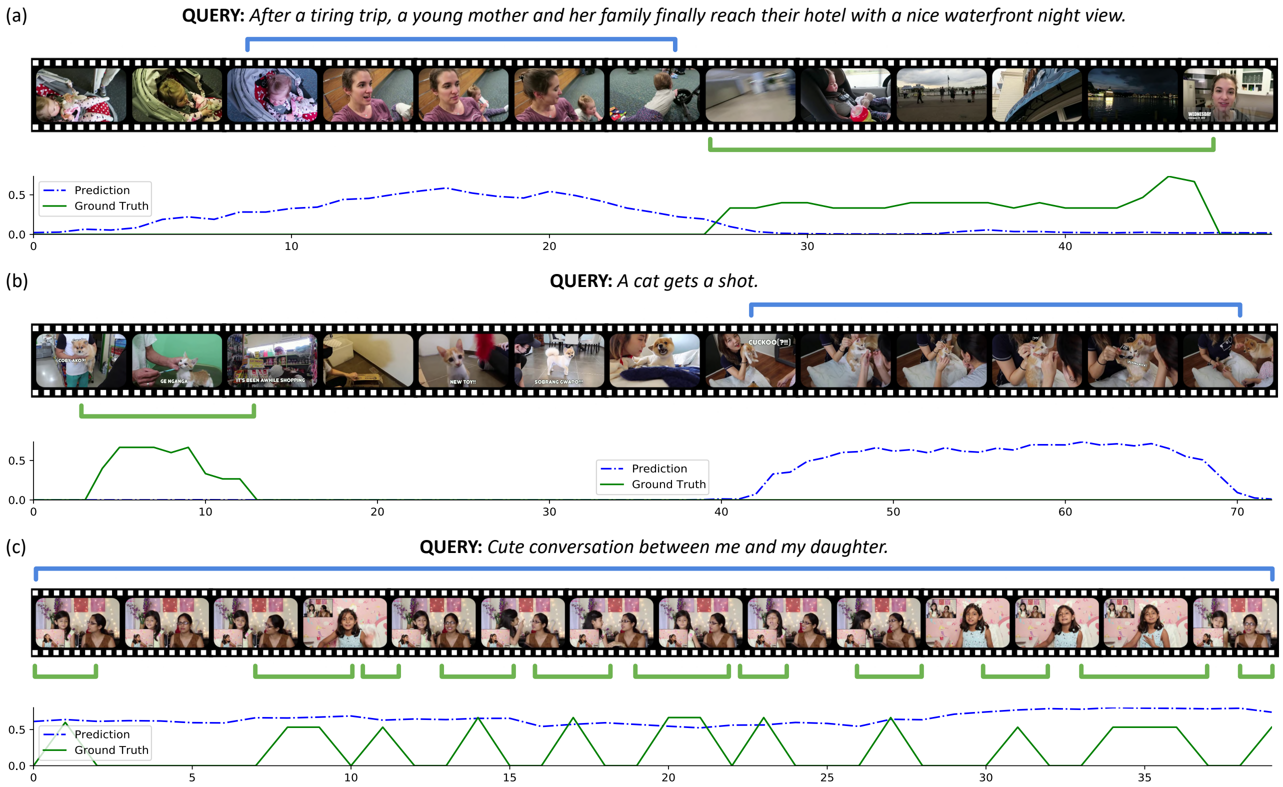
(b) **QUERY:** *A dog circling in a lawn.*

(c) **QUERY:** *A girl looks happy to be laying next to her fluffy white cat.*

(d) **QUERY:** *Black and white photo of a man indicating we are listening to his audio.*

(e) **QUERY:** *Group of friends listening to music and enjoying in car.*

Figure 1. Visualization results on QVHighlights `val` split.



(a) **QUERY:** *After a tiring trip, a young mother and her family finally reach their hotel with a nice waterfront night view.*

(b) **QUERY:** *A cat gets a shot.*

(c) **QUERY:** *Cute conversation between me and my daughter.*

Figure 2. Failure cases on QVHighlights `val` split.