Appendix

A. Additional Details about Bidirectional Modality Random Walk Algorithm.

As stated in the manuscript, we propose a Bidirectional Modality Random Walk algorithm (BMRW) to exploit the power of the fine-grained multi-modal interaction. In this section, we provide in-depth illustration on it. First, we define the linguistic feature $\mathcal{F}_p^{(0)}$ and visual feature $\mathcal{F'}_v^{(0)}$ as initial variables at 0-th iteration:

$$\begin{cases} \mathcal{F}_p^{(0)} = \mathcal{F}_p \\ \mathcal{F'}_v^{(0)} = \mathcal{F'}_v \\ \mathcal{Z} = \lambda_z \mathcal{F'}_v^{(0)} (\mathcal{F}_p^{(0)})^\top, \end{cases} \tag{19}$$

where $\mathcal{Z}$ is the affinity between two modalities. Then we propagate the semantics between modalities in two iterative formulas:

$$\begin{cases} \mathcal{F}_p^{(t)} = \omega \text{Norm1}(\mathcal{Z})^\top \mathcal{F'}_v^{(t-1)} + (1-\omega)\mathcal{F}_p^{(0)} \\ \mathcal{F'}_v^{(t)} = \omega \mathcal{Z} \mathcal{F}_p^{(t)} + (1-\omega)\mathcal{F'}_v^{(0)}, \quad \omega \in (0,1). \end{cases} \tag{20}$$

Integrating Eq. (20) we can get:

$$\begin{aligned} \mathcal{F'}_v^{(t)} &= (\omega^2 Z Norm1(Z)^\top)\mathcal{F'}_v^{(t-1)} \\ &\quad + \omega(1-\omega)\mathcal{Z}\mathcal{F}_p^{(0)} + (1-\omega)\mathcal{F'}_v^{(0)} \\ &= (\omega^2 Z Norm1(Z)^\top)\mathcal{F'}_v^{(t-1)} \\ &\quad + (1-\omega)(\omega Z \mathcal{F}_p^{(0)} + \mathcal{F'}_v^{(0)}). \end{aligned} \tag{21}$$

We substitute $ZNorm1(Z)^\top$ as A and expand Eq. (21) from t-th to 0-th iteration:

$$\mathcal{F'}_v^{(t)} = (\omega^2 A)^t \mathcal{F'}_v^{(0)} + (1-\omega)\sum_{i=0}^{t-1}(\omega^2 A)^i(\omega \mathcal{Z}\mathcal{F}_p^{(0)} + \mathcal{F'}_v^{(0)}), \tag{22}$$

Considering the potential risk of unexpected gradient or expensive computation cost, we use an approximate optimal solution based on Neumann Series [30]:

$$\lim_{t\to\infty} \sum_{i=0}^{t-1} (\omega^2 A)^i = (I - \omega^2 A)^{-1}. \tag{23}$$

Accordingly, when $t \to \infty$ in Eq. (22), the comprehensive visual representation is generated as:

$$\mathcal{F'}_v^{(\infty)} = (1-\omega)(I - \omega^2 A)^{-1}(\omega \mathcal{Z}\mathcal{F}_p^{(0)} + \mathcal{F'}_v^{(0)}). \tag{24}$$

B. Additional Details on Experiment Settings

B.1. Datasets

QVHighlight is the most popular publicized dataset which consists of over 10,000 videos with human-written free-form text descriptions for moment retrieval and highlight detection. Charades-STA and TACoS are both for moment retrieval where Charades-STA comprises 16,128 query-moment pairs for indoor activities and TACoS contains 127 annotated videos from cooking scenarios. TVSum and YouTube Highlights cater for highlight detection. Each of which includes 10 domains with 5 videos and 6 domains with 433 videos respectively.

Method	R1@0.5	R1@0.7
SAP [4]	27.42	13.36
SM-RL [49]	24.36	11.17
MAN [61]	41.24	20.54
2D-TAN [64]	40.94	22.85
QD-DETR [31]	52.77	31.13
UVCOM	**54.57**	**34.13**
UMT† [27]	48.31	29.25
QD-DETR† [31]	55.51	34.17
UVCOM†	**56.69**	**34.76**

Table 9. **MR results on Charades-STA Test Split**. The pre-extracted features are from VGG, GLOVE Embeddings and PANN. † denotes using audio modality.

B.2. Metrics

Recall@1 with IoU thresholds 0.5 and 0.7, mean average precision (mAP) with IoU thresholds 0.5 and 0.75 as well as average mAP over 0.5:0.05:0.95 are for MR, while mAP and HIT@1 are used for HD. HIT@1 is computed through the hit ratio of the clip with the highest score. For Charades-STA and TACoS, we report the result of Recall@1 with IoU thresholds 0.5 and 0.7. For YouTube Highlights and TVSum, we follow [27] and adopt the metrics of mAP and Top-5 mAP.

B.3. Feature Representations

The pre-extracted visual and text features from SlowFast [6] and CLIP [36] are used on all datasets. Notably, on Charades-STA and YouTube Highlights, we additionally extract the features from official VGG [41] as well as GLOVE [35] embeddings and commonly used I3D [3], respectively for further comparisons. Besides, we leverage PANN [19] model to encode audio features for experiments with extra audio modality learning.

B.4. Training Details

Elaborate parameter settings for each benchmark are summarized in Tab. 10. For more details, all experiments are implemented in PyTorch with one 24GB RTX3090. The overall aggregation for DBIA is performed in 5 iterations. In addition, the hidden dimension of transformer is 256 for all experiments.

Dataset	Feature	Lr	Epoch	Bs	Lr drop	n_v	n_t	λ_{gIoU}	λ_{L1}	λ_{HD}	λ_{hard}	λ_{cta}	λ_{vld}
QVHighlights	SF+C	1e-4	200	32	100	30	5	1	10	1	1	0.5	0.5
Charades-STA	SF+C	1e-4	200	8	80	30	5	1	10	1	1	0.5	0.5
Charades-STA	VGG	1e-4	200	8	-	30	5	1	10	1	1	1.5	0.5
TaCoS	SF+C	1e-4	200	32	100	30	5	1	10	1	1	0.5	0.5
TVSum	I3D	1e-4	2000	4	-	30	5	1	10	1	1	Tab. 15	Tab. 15
YouTubeHL	SF+C	1e-4	2000	4	1000	30	2	-	-	1	1	Tab. 16	Tab. 16
YouTubeHL	I3D	1e-4	2000	4	1000	30	2	-	-	1	1	Tab. 17	Tab. 17

Table 10. **Training details.** We provide elaborate training details on each dataset. Lr denotes learning rate; Bs denotes batch size; Lr drop denotes the drop of learning rate at the specific epoch. n_v and n_t denote the number of Guassians in DBIA module.

n_v	n_t	MR			HD	
		R1 @0.5	R1 @0.7	mAP Avg.	mAP	HIT@1
20	3	62.58	48.45	42.44	39.58	61.87
25	4	64.45	50.06	42.95	40.0	**64.9**
25	5	64.26	50.06	43.48	39.51	62.06
30	**5**	**65.10**	**51.81**	**45.79**	**40.03**	63.29
30	6	62.19	48.52	43.53	39.9	64.32

Table 11. **Results of different numbers of Gaussian for video and text.**

λ_{cta}	λ_{vld}	MR			HD	
		R1 @0.5	R1 @0.7	mAP Avg.	mAP	HIT@1
0.3	0.7	63.61	48.19	43.68	39.88	63.68
0.5	1.0	64.84	50.13	44.69	40.02	**64.13**
0.5	**0.5**	**65.10**	**51.81**	**45.79**	**40.03**	63.29
1.5	0.7	64.13	49.81	43.97	39.96	63.48

Table 12. **Results of the different hyper-parameters in Multi-Aspect Contrastive Learning.**

Iterations	MR			HD	
	R1 @0.5	R1 @0.7	mAP Avg.	mAP	HIT@1
3	63.55	48.52	43.19	40.03	64.26
4	63.48	50.0	44.09	39.77	**64.71**
5	**65.10**	**51.81**	**45.79**	**40.03**	63.29
6	62.65	47.94	43.09	39.44	62.77

Table 13. **Results of different iterations of EM Attention.**

Layer Num.	MR			HD	
	R1 @0.5	R1 @0.7	mAP Avg.	mAP	HIT@1
2	61.74	46.52	41.18	39.81	63..1
3	**65.10**	**51.81**	**45.79**	**40.03**	**63.29**
4	63.68	49.61	43.8	39.63	62.65
5	63.55	49.16	43.73	39.56	62.77

Table 14. **Results of different numbers of encoder and decoder layers.**

C. Additional Experiments

C.1. Result on Charades-STA

We also present comparisons on Charades-STA [7] with existing methods in Tab. 9. As observed, our UVCOM achieve the new state-of-the-art performance under different settings, which further validates the rationality of our design in local perception enhancement for MR.

C.2. Additional Ablation Studies

We conduct additional analysis experiments on the val split of QVHighlights benchmark.

Number of Gaussian. The degree of the intra-modality aggregation is determined by the number of Gaussian. Specifically, we conduct ablation studies on two parameters n_v and n_t where each denotes the number of condense visual and linguistic outputs in Tab. 11. We find that the performance tends to boost as the number of Gaussians increases for which the smaller one may lead to inefficiency for content clustering in video or text. However, the large value of Guassians degrades the performance. We assume that the larger one introduces redundancy, which hinders the converge process.

Multi-Aspect Contrastive Learning Coefficients. For multi-aspect contrastive learning, $\mathcal{L}_{cta}$, $\mathcal{L}_{vld}$ are coefficients for clip-text alignment and video-linguist discrimination loss respectively. We report the ablation results in Tab. 12. As observed, the appropriate setting of coefficient helps decently solidify the local relation modeling and global knowledge integration, thereby facilitating the comprehensive understanding.

Aggregation Iteration. The aggregation iteration controls the quality of multi-grained feature generation, *i.e.*, moment-level and phrase-level features. It can be seen in Tab. 13 that the insufficient or excessive iterations both lead

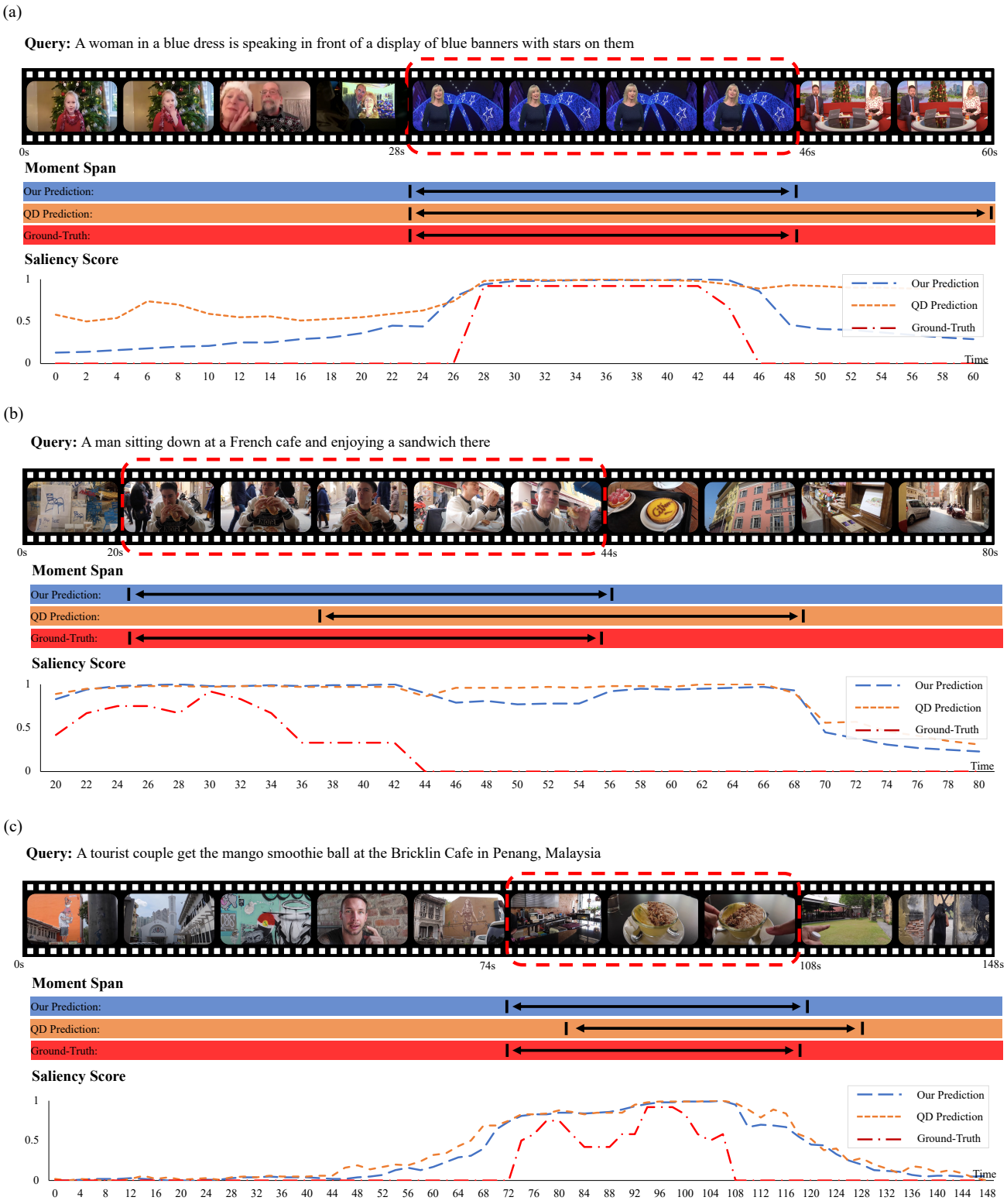
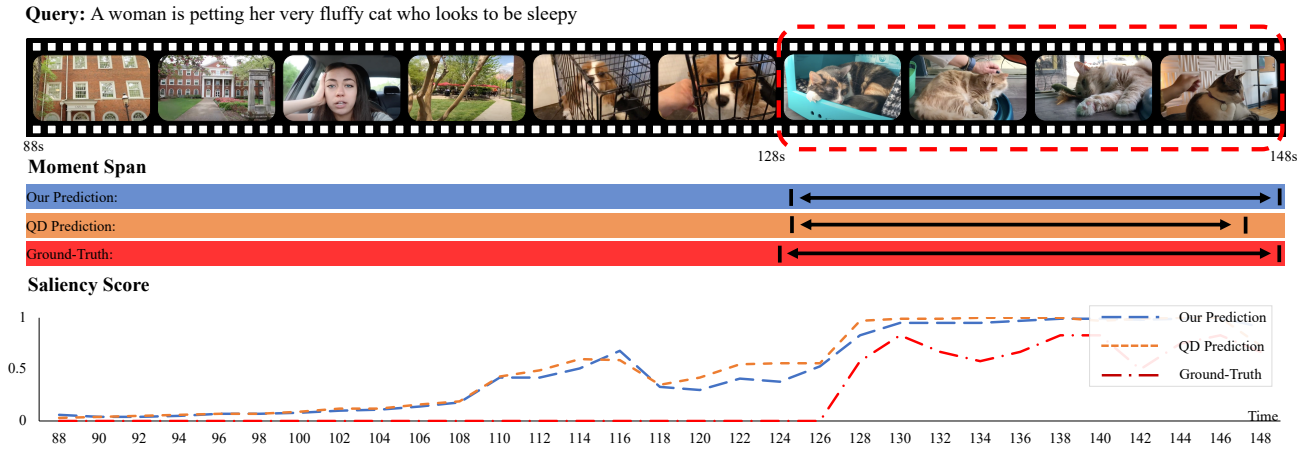
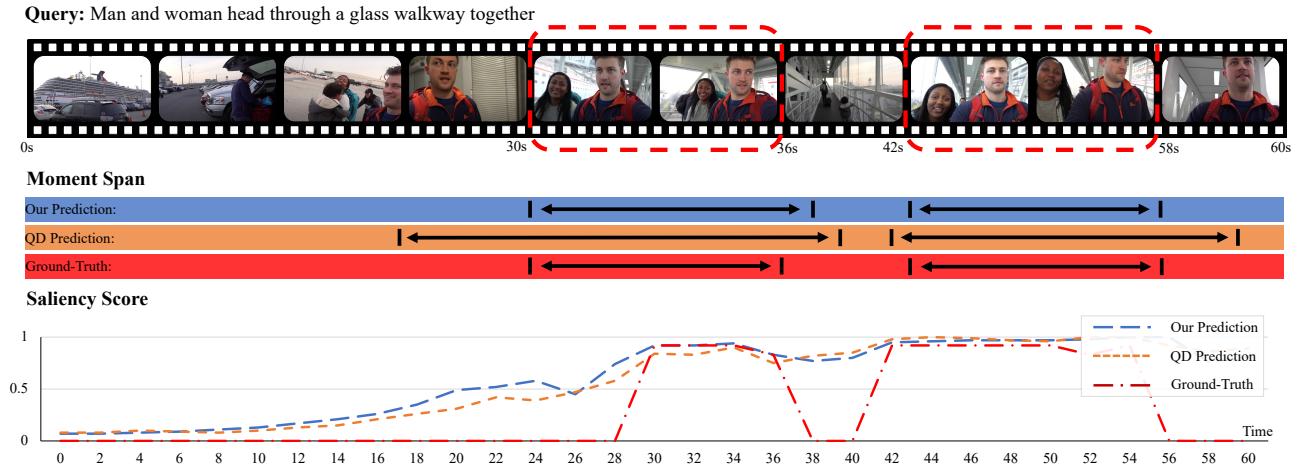


Figure 7. **Visualization comparison on MR and HD.** QD indicates previous state-of-the-art method QD-DETR [31]

(e)



(f)



(g)

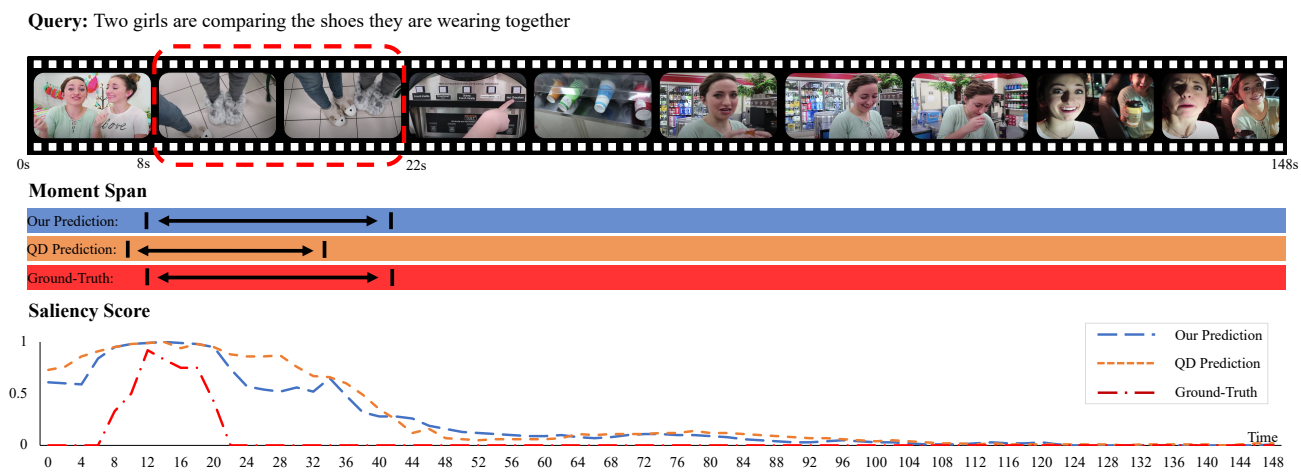


Figure 8. **Visualization comparison on MR and HD.** QD indicates previous state-of-the-art method QD-DETR [31]

Domain	VT	VU	GA	MS	PK	PR	FM	BK	BT	DS
λ_{cta}	0.4	0.75	0.5	0.1	0.3	1.0	0.25	0.5	1	0.25
λ_{vld}	0.2	0.75	0.5	0.1	0.3	0	0.25	0.5	0	0.25

Table 15. λ_{cta} **and** λ_{vld} **for TVSum.**

Domain	Dog	Gym.	Park.	Ska.	Ski.	Surf.
λ_{cta}	0.5	0.8	0	0.5	0	0.5
λ_{vld}	1	0.8	1.5	1	1.5	1

Table 16. λ_{cta} **and** λ_{vld} **for YoutubeHL using SF+C feature.**

Domain	Dog	Gym.	Park.	Ska.	Ski.	Surf.
λ_{cta}	0.5	0.8	1.0	0.5	1.0	0.5
λ_{vld}	0.5	0.8	1.5	1	1.5	0.5

Table 17. λ_{cta} **and** λ_{vld} **for YoutubeHL using I3D feature.**

to the decreased performance due to the incomplete contextual information fusion or over exaggeration on similar contents.

Layers. To investigate the impact on the different encoder and decoder layers, we provide the performance variation in Tab. 14. The results depict that the increased layers bring significant improvement. However, the performance saturates when the number of layers reaches 5. This can be attributed to the noises accumulation caused by redundant interaction.

C.3. Visualizations.

Fig. 7 and Fig. 8 display additional qualitative comparisons between our UVCOM and a previous state-of-the-art method, which shows our consistent performance on Moment Retrieval and Highlight Detection.