Cross-modal Multiscale Difference-aware Network for Joint Moment Retrieval and Highlight Detection

Mingyao Zhou[1,2,3], Wenjing Chen[4], Hao Sun[1,2,3†], Wei Xie[1,2,3†]

[1]Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning, Central China Normal University, Wuhan, China

[2]School of Computer, Central China Normal University, Wuhan, China

[3]National Language Resources Monitoring and Research Center for Network Media, Central China Normal University, Wuhan, China

[4]School of Computer Science, Hubei University of Technology, Wuhan, China

[†]Corresponding authors: haosun@mail.ccnu.edu.cn, XW@mail.ccnu.edu.cn

Motivation

Although existing methods [1-2] for joint moment retrieval and highlight detection achieve impressive performance, they still face some problems.

- **Semantic gaps across different modalities.**
- **Smooth transitions among diverse events.**
- **Various durations of different query-relevant moments and highlights.**

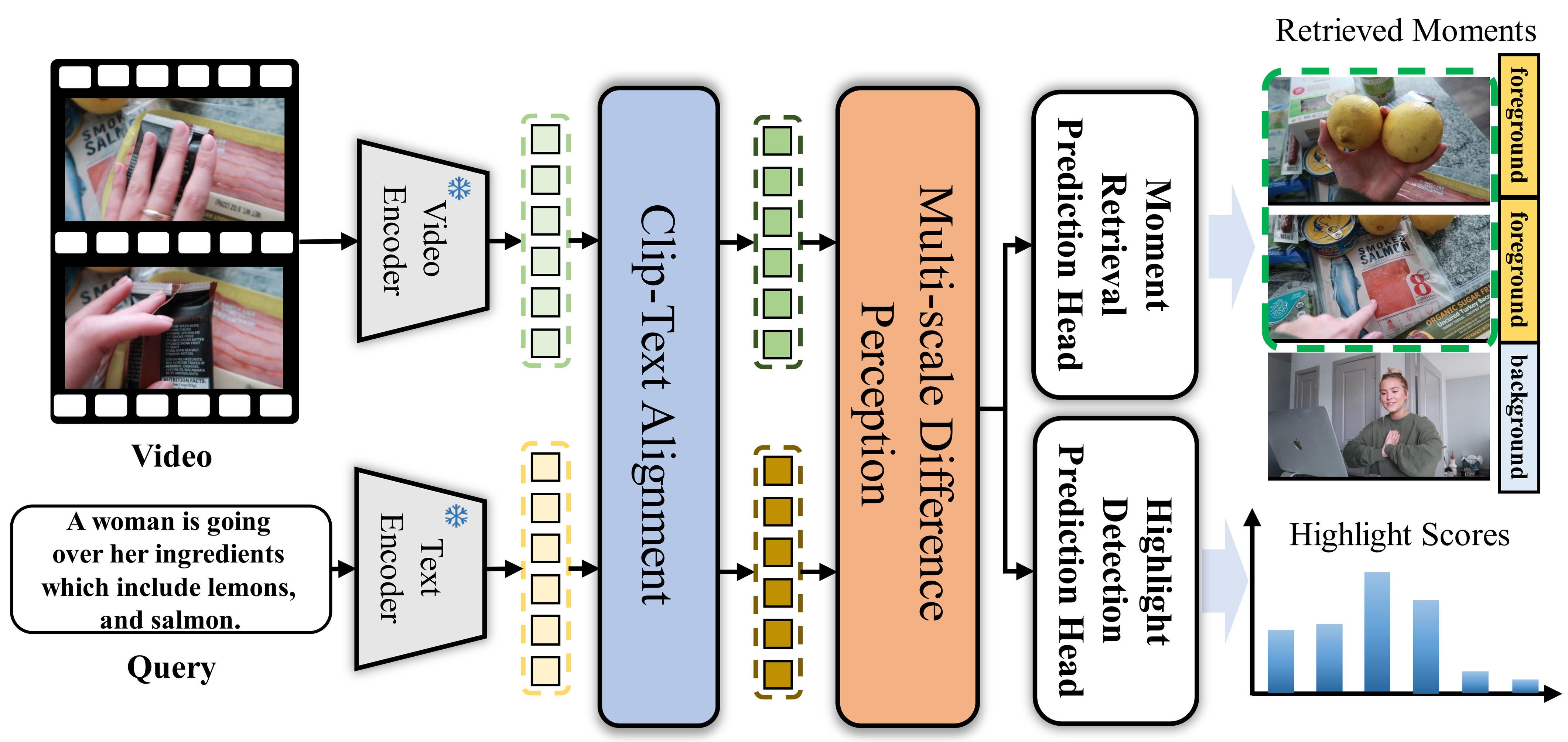
Therefore, a Cross-modal Multiscale Difference-aware Network is proposed.

Contribution

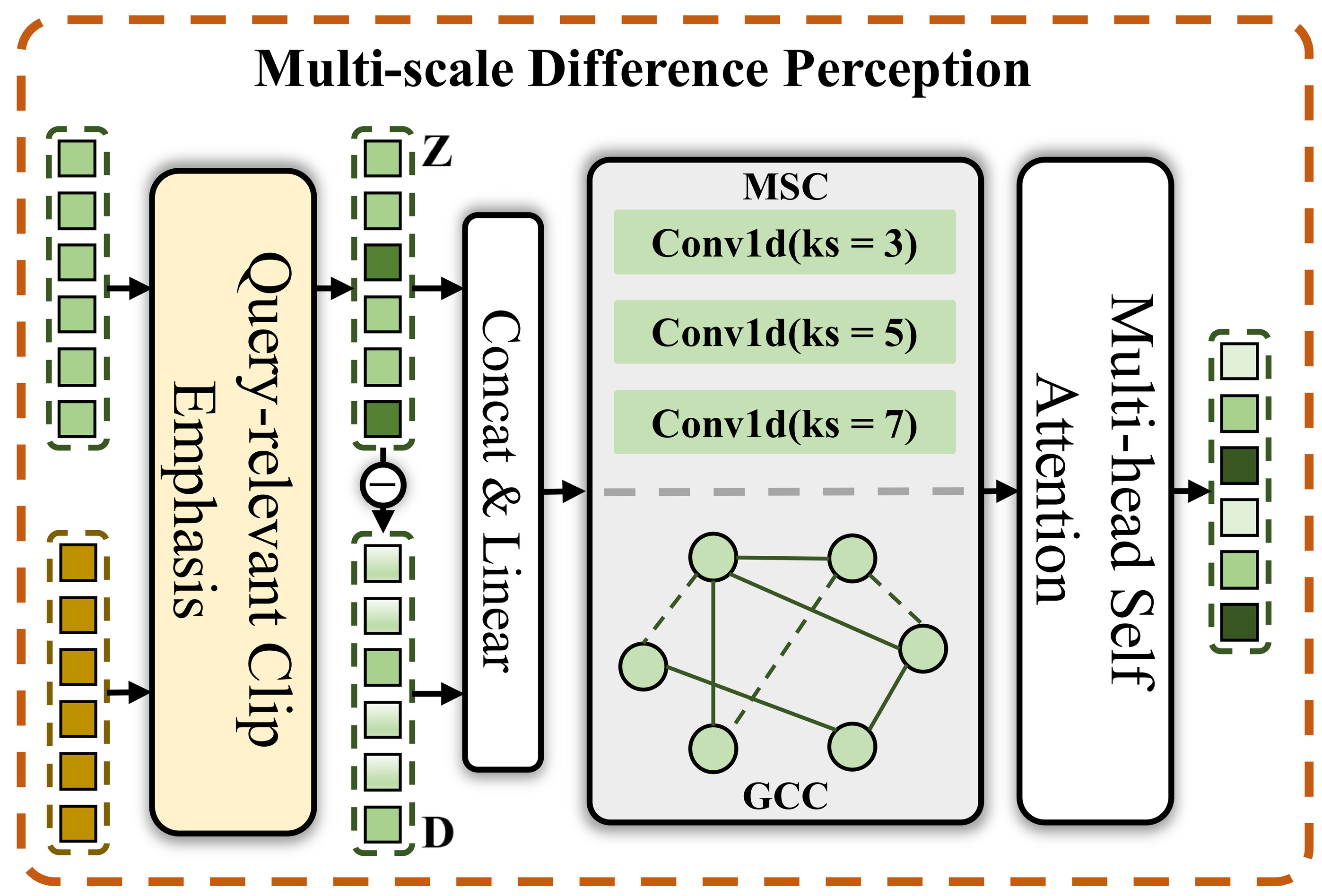
The contributions of this paper are as follows:

- ✓ Build a clip-text alignment module to alleviate the modal gaps.
- ✓ Propose a multi-scale difference perception module to fully integrate the differential information of adjacent clips and obtain joint representations through multi-scale modeling.
- ✓ A large number of experiments prove the effectiveness of the proposed method.

Methodology



a) An overview of the proposed cross-modal multiscale difference-aware network (CMDNet).



b) Design details of multi-scale difference perception module.

- We start by extracting visual and textual features using pre-trained encoders.
- Then, we create a cross-modal alignment module inspired by CLIP [3] to bridge the semantic gaps between text and video. This module uses **inner modality constraints** to refine video representations and **intermodal constraints** to align text and visual features.
- To alleviate issues posed by smooth transitions in video events, we introduce a multi-scale perception module that highlights query-related features and **incorporates differential information** between adjacent clip features.
- To enhance diversity in target moment durations, we utilize multi-scale convolution and graph convolution components. These components **capture temporal dependencies at different scales** and model global dependencies in the video.

Results & Discussion

Method	Moment Retrieval					HD	
	R1		mAP			≥ Very Good	
	@0.5	@0.7	@0.5	@0.75	Avg.	mAP	HIT@1
XML+ [2]	46.69	33.46	47.89	34.67	34.90	35.38	55.06
MDETR [2]	52.89	33.02	54.82	29.40	30.73	35.69	55.60
MHDETR [6]	60.05	42.48	60.75	38.13	38.38	38.22	60.51
QDDETR [4]	62.40	44.98	62.62	39.88	39.86	38.64	**62.40**
UniVTG [15]	58.86	40.86	57.60	35.59	35.47	38.20	60.96
CMDNet	**62.52**	**46.69**	**63.63**	**43.44**	**42.89**	**39.80**	62.26
UMT [3]	56.23	41.18	53.38	37.01	36.12	38.18	59.99
MIM [5]	59.99	41.50	55.85	36.84	36.45	38.96	62.39
QDDETR [4]	63.06	45.10	63.04	40.10	40.19	39.04	62.87
CMDNet	**63.62**	**47.28**	**63.89**	**44.12**	**43.23**	**40.06**	**63.16**

Tab.1: Experimental results on the QVHighlights test set. The lower half of the table represents the introduction of audio information.

Method	R1@0.5	R1@0.7	Method	R1@0.5	R1@0.7
2DTAN [18]	40.94	22.85	VSLNet [17]	47.31	30.19
FVMR [19]	42.36	24.14	QDDETR [4]	50.67	31.02
UMT† [3]	48.31	29.25	CMDNet	**56.24**	**35.16**
QDDETR [4]	52.77	31.13	MDETR [2]	53.63	31.37
QDDETR† [4]	**55.51**	**34.17**	QDDETR [4]	57.31	32.55
CMDNet	53.33	33.47	UniVTG [15]	58.01	35.65
CMDNet†	54.97	33.52	CMDNet	**58.55**	**36.16**

Tab.2: Results on Charades-STA. '†' denotes introducing extra audio information.

- Judging from the experimental results presented here, Our approach outperforms others.
- But when we directly splice audio information without alignment modeling, performance suffered.
- In the future, we'll focus on leveraging audio's semantic information for MR and HD tasks, effectively.

References

1. Lei, J.; Berg, T. L.; and Bansal, M. 2021. Detecting Moments and Highlights in Videos via Natural Language Queries. In NeurIPS, 11846–11858.
2. Moon, W.; Hyun, S.; Park, S.; Park, D.; and Heo, J.-P. 2023. Query-dependent video representation for moment retrieval and highlight detection. In IEEE CVPR, 23023–23033.
3. Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning transferable visual models from natural language supervision. In ICML, 139:8748–8763.