MomentDiff: Generative Video Moment Retrieval from Random to Real

Pandeng Li,[1] Chen-Wei Xie,[2] Hongtao Xie,[1*] Liming Zhao,[2] Lei Zhang,[1]
Yun Zheng,[2] Deli Zhao,[2] Yongdong Zhang[1]

[1] University of Science and Technology of China
[2] Alibaba Group
lpd@mail.ustc.edu.cn

Video Moment Retrieval

Text query: "*Kids checking out their goodies and chocolates during Easter.*."

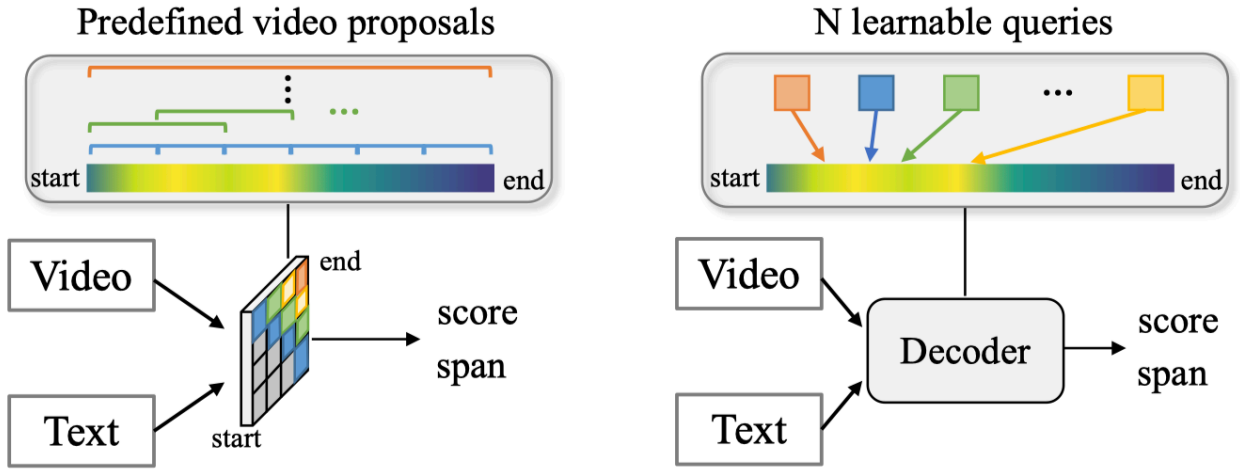
Video:

[0s, 60s]

Video moment retrieval aims to identify the **moment boundaries** (the start and end time) within a given video that best semantically correspond to the text query.

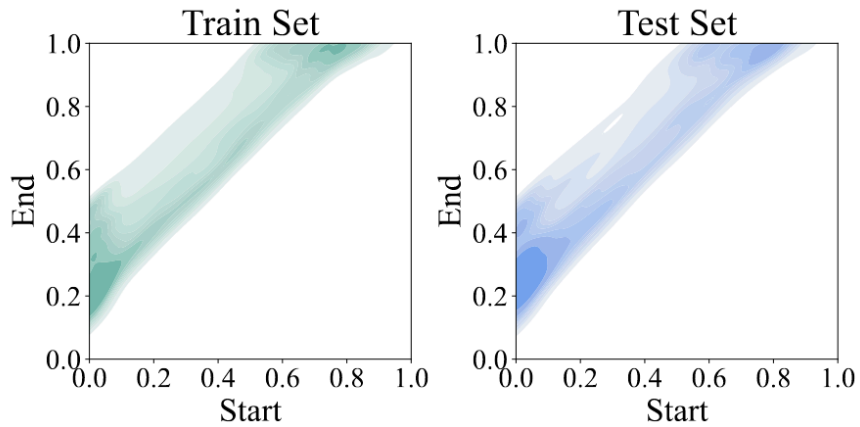
Motivation

- **Temporal location bias**



(a) Dense: 2DTAN

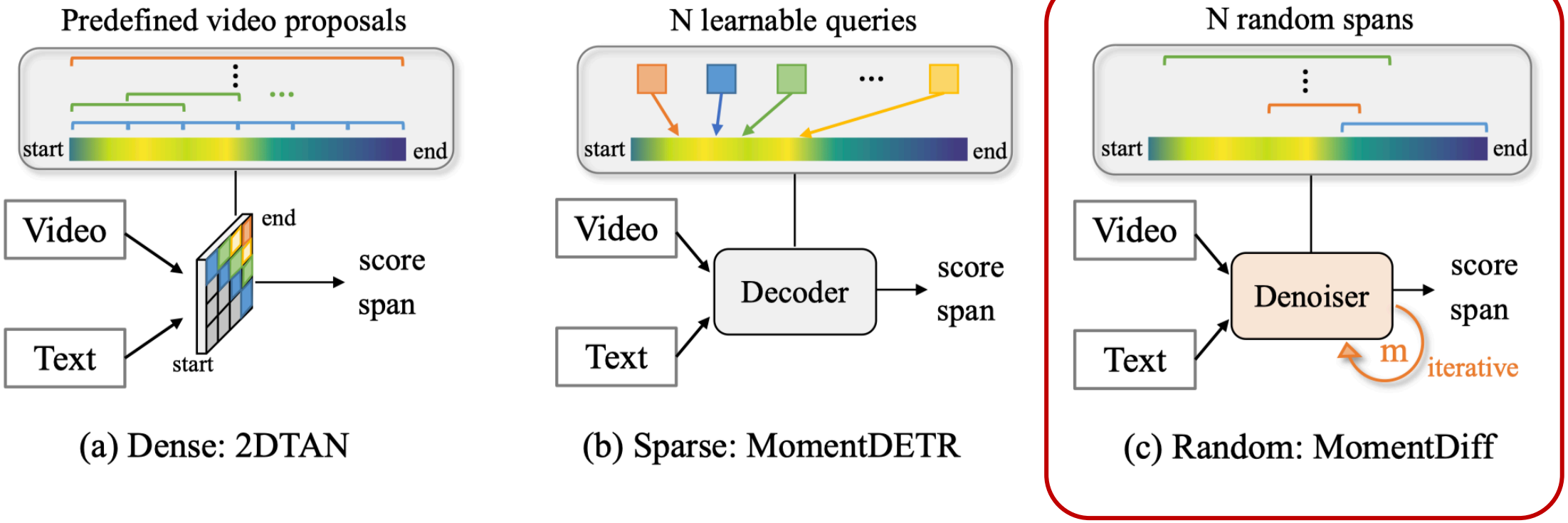
(b) Sparse: MomentDETR



Moment distributions for Charades-STA [1]

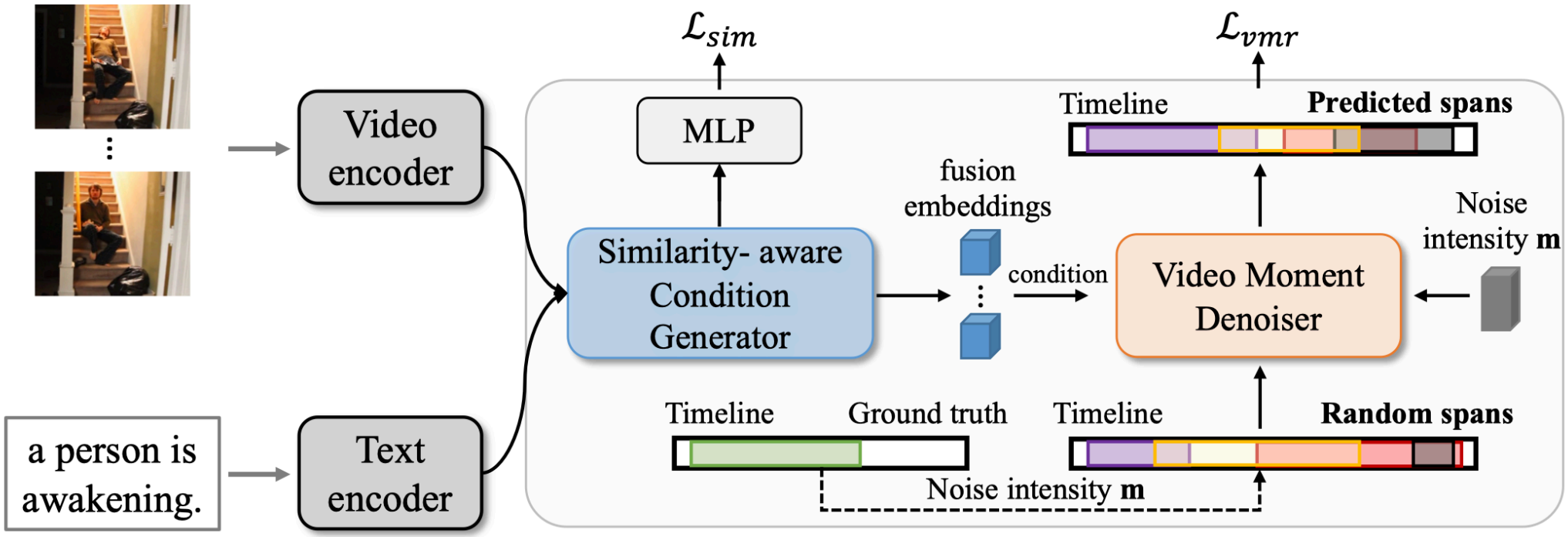
[1] Otani, M., Y. Nakashima, E. Rahtu, et al. Uncovering hidden challenges in query-based video moment retrieval. In BMVC. 2020.

Our idea


Predefined video proposals
start
end
Video
Text
end
start
score
span
(a) Dense: 2DTAN
N learnable queries
start
end
Video
Text
Decoder
score
span
(b) Sparse: MomentDETR
N random spans
start
end
Video
Text
Denoiser
score
span
m
iterative
(c) Random: MomentDiff

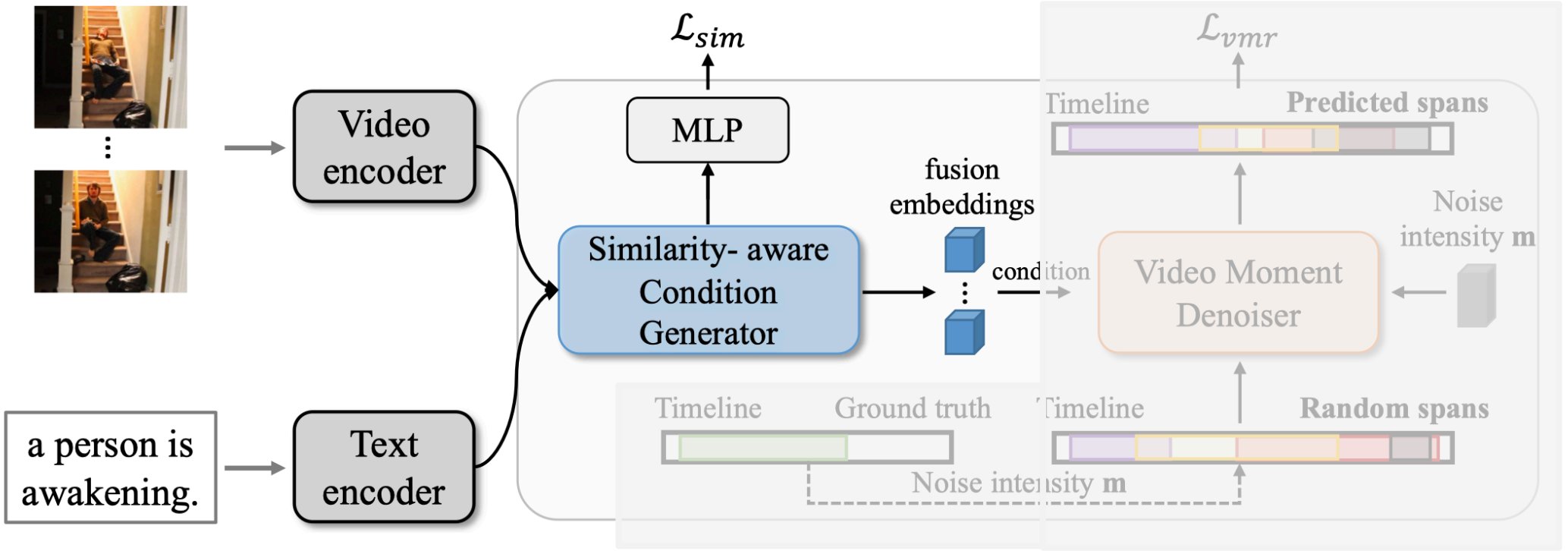
Method

- **Framework**



Method

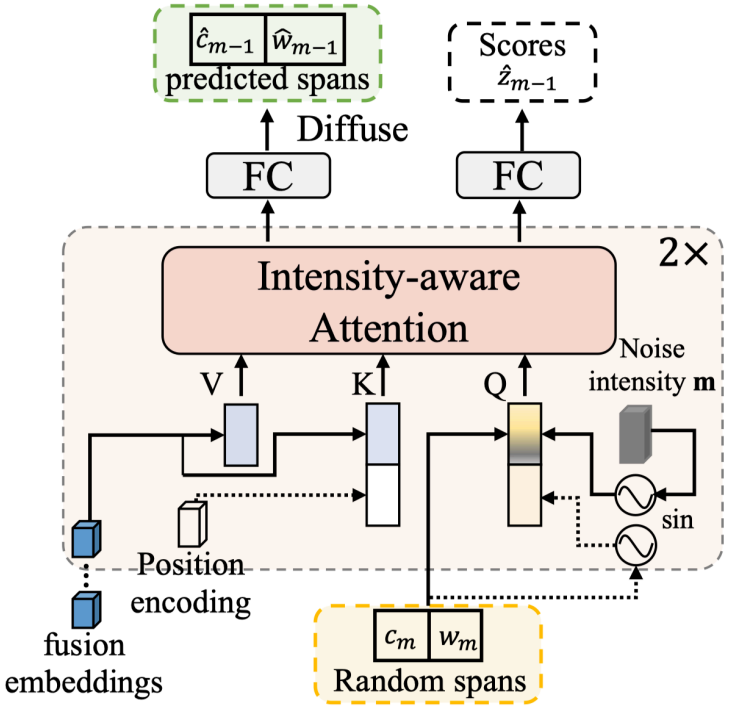
- **Similarity-aware Condition Generator (SCG)**



$$\mathcal{L}_{sim} = -\frac{1}{N_v}\sum_{i=1} \boldsymbol{y}_i * log(\boldsymbol{s}_i) + (1-\boldsymbol{y}_i) * log(1-\boldsymbol{s}_i) + \frac{1}{N_s}\sum_{j=1} \max\left(0, \beta + \boldsymbol{s}_{n_j} - \boldsymbol{s}_{p_j}\right)$$

Method

- **Video Moment Denoiser (VMD)**



1. Span normalization
2. Span embedding
3. Intensity-aware attention
4. Denoising training

$$\mathcal{L}_{\text{vmr}}\left(\boldsymbol{x}_0, f_\theta(\boldsymbol{x}_m, m, \boldsymbol{F})\right) = \lambda_{\text{L1}} \left\|\boldsymbol{x}_0 - \hat{\boldsymbol{x}}_{m-1}\right\| + \lambda_{\text{iou}} \mathcal{L}_{\text{iou}}\left(\boldsymbol{x}_0, \hat{\boldsymbol{x}}_{m-1}\right) + \lambda_{\text{ce}} \mathcal{L}_{ce}(\hat{\boldsymbol{z}}_{m-1})$$

Experiment

Charades-STA

Table 1: Performance comparisons (%) on the Charades-STA dataset. "⋆" denotes that we re-implement the method under the same training scheme. "A" stands for using audio data.

Method	Type	Charades-STA				
		R1@0.5	R1@0.7	MAP@0.5	MAP@0.75	MAP_{avg}
MAN [32]	VGG, Glove	41.21	20.54	-	-	-
RaNet* [74]	VGG, Glove	42.91	25.82	53.28	24.41	28.55
2DTAN* [26]	VGG, Glove	41.34	23.91	54.68	24.15	29.26
DORi [77]	VGG, Glove	43.47	26.37	-	-	-
CBLN [43]	VGG, Glove	47.94	28.22	-	-	-
DCM [39]	VGG, Glove	47.80	28.00	-	-	-
MMN* [27]	VGG, Glove	46.93	27.07	58.85	28.16	31.58
MomentDETR* [36]	VGG, Glove	50.54	28.01	57.39	25.62	29.87
MomentDiff	VGG, Glove	**51.94**	**28.25**	**59.86**	**29.11**	**31.66**
UMT [37]	VGG+A, Glove	48.44	29.76	58.03	27.46	30.37
MomentDiff	VGG+A, Glove	**52.62**	**29.93**	**60.69**	**29.74**	**31.81**
DEBUG [78]	C3D, Glove	37.39	17.69	-	-	-
LPNet [35]	C3D, Glove	40.94	21.13	-	-	-
VSLNet* [29]	C3D, Glove	48.67	30.33	56.88	25.79	30.16
MomentDETR* [36]	C3D, Glove	50.49	29.95	56.27	26.08	29.92
MomentDiff	C3D, Glove	**53.79**	**30.18**	**59.32**	**29.85**	**31.89**
MomentDETR* [36]	SF+C, C	53.22	30.87	58.86	26.43	30.43
MomentDiff	SF+C, Glove	55.42	32.17	60.93	32.47	32.59
MomentDiff	SF+C, C	**55.57**	**32.42**	**61.07**	**32.51**	**32.85**

Experiment

QVHighlights

Table 2: Performance comparisons (%) on QVHighlights with SF+C video features and CLIP text features. "⋆" de-notes that we re-implement the method with only segment moment labels. "†" stands for using audio data. MDE is the abbreviation of MomentDETR [36].

Method	QVHighlights				
	R1@0.5	R1@0.7	MAP@0.5	MAP@0.75	MAP_{avg}
MCN [24]	11.41	2.72	24.94	8.22	10.67
CAL [79]	25.49	11.54	23.40	7.65	9.89
XML [80]	41.83	30.35	44.63	31.73	32.14
XML+ [80]	46.69	33.46	47.89	34.67	34.90
MDE⋆ [36]	53.56	34.09	53.97	28.65	29.39
MomentDiff	**57.42**	**39.66**	**54.02**	**35.73**	**35.95**
UMT⋆† [37]	56.26	40.31	52.77	36.82	35.79
MomentDiff†	**58.21**	**41.48**	**54.57**	**37.21**	**36.84**

TACoS

Table 3: Performance comparisons (%) with C3D features to encode videos. MDE is the abbreviation of MomentDETR [36].

Method	TACoS		
	R1@0.1	R1@0.3	R1@0.5
CTRL [22]	24.32	18.32	13.30
SCDM [31]	-	26.11	21.17
DRN [30]	-	-	23.17
DCL [40]	49.36	38.84	29.07
CBLN [43]	49.16	38.98	27.65
FVMR [41]	53.12	41.48	29.12
RaNet [74]	-	43.34	33.54
MDE⋆ [36]	41.16	32.21	20.55
MMN⋆ [27]	51.39	39.24	26.17
MomentDiff	**56.81**	**44.78**	**33.68**

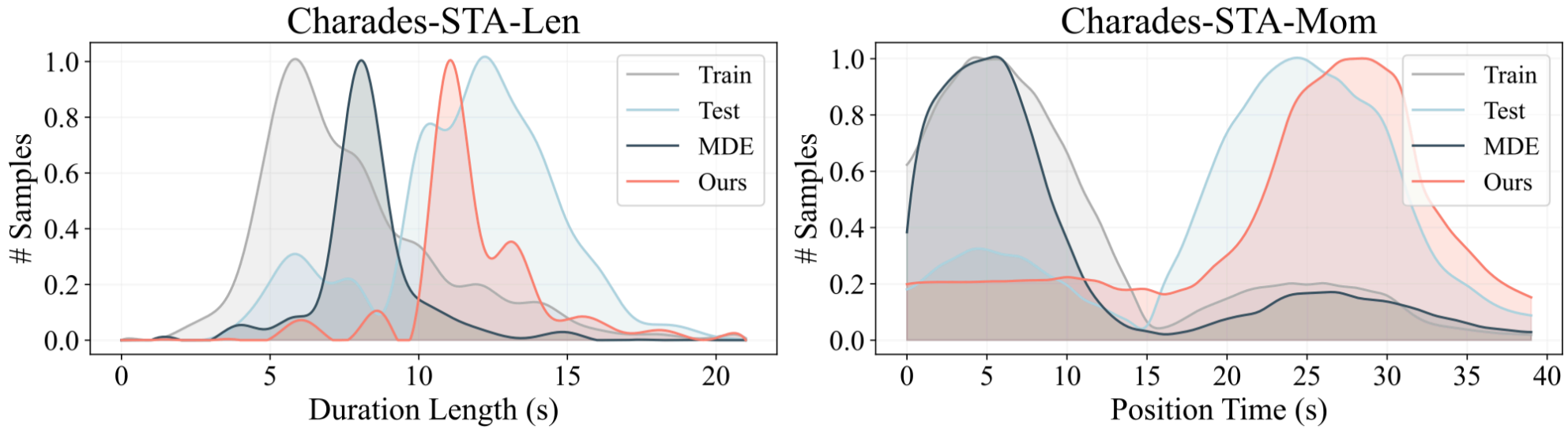
Experiment

R1@n and MAP results on public OOD datasets

Method	Charades-CD				ActivityNet-CD			
	R1@0.3	R1@0.5	R1@0.7	MAP	R1@0.3	R1@0.5	R1@0.7	MAP
2DTAN	49.71	28.95	12.78	12.60	40.04	22.07	10.29	12.77
MMN	55.91	34.56	15.84	15.73	44.13	24.69	12.22	15.06
MomentDETR	57.34	41.18	19.31	18.95	39.98	21.30	10.58	12.19
MomentDiff	**67.73**	**47.17**	**22.98**	**22.76**	**45.54**	**26.96**	**13.69**	**16.38**

Experiment

Statistical distributions on our anti-bias datasets



Temporal moment $\hat{x}_0 = (\hat{c}_0, \hat{w}_0)$

$\hat{c}_0$: The center time of moment

$\hat{w}_0$: The duration length of moment

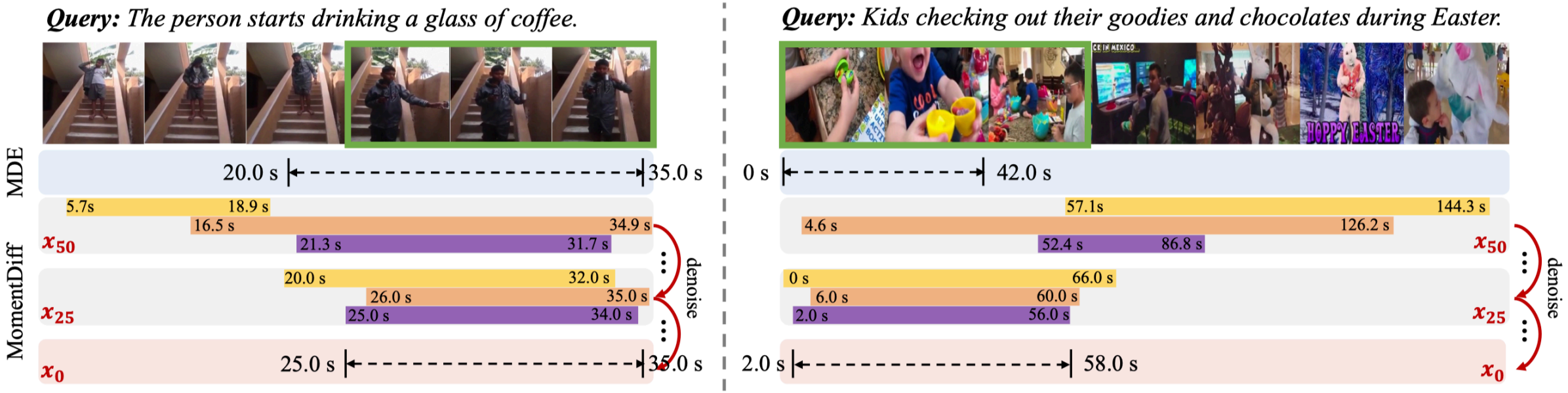
Experiment

R1@n and MAP results on anti-bias datasets

Method	Charades-STA-Len				Charades-STA-Mom			
	R1@0.3	R1@0.5	R1@0.7	MAP	R1@0.3	R1@0.5	R1@0.7	MAP
2DTAN	39.68	28.68	17.72	22.79	27.81	20.44	10.84	17.23
MMN	43.58	34.31	19.94	26.85	33.58	27.20	14.12	19.18
MomentDETR	42.73	34.39	16.12	24.02	29.94	22.16	11.56	18.66
MomentDiff	**51.25**	**38.32**	**23.38**	**28.19**	**48.39**	**33.59**	**15.71**	**21.37**

Experiment

Visualization of the diffusion process



MomentDiff: Generative Video Moment Retrieval from Random to Real

Thank you!

Github Code

Homepage