–Supplementary Material– UniMD: Towards Unifying Moment Retrieval and Temporal Action Detection

Yingsen Zeng, Yujie Zhong✉, Chengjian Feng, and Lin Ma

Meituan Inc.

1 Appendix

In the appendix, we describe (1) more details about three paired datasets (Section A), (2) details about implementation (Section B), (3) additional ablation experiments (Section C), and (4) qualitative results (Section D). For sections, figures, tables, and equations, we use numbers (*e.g.* Sec. 1) to refer to the main paper and capital letters (*e.g.* Sec. A) to refer to this appendix.

A Details About Datasets

Statistical analysis of three paired datasets. (1) **Ego4D** [3] is a large-scale first-person video dataset that consists of ∼3000 hours of egocentric videos. The TAD task of Ego4D includes 110 action categories. Specifically, the TAD task has an average of 9.2 instances per video across 2k videos, while the MR task has an average of 10.2 instances per video across 1.3k videos. (2) **Charades and Charades-STA.** Charades [7] is a large-scale dataset that consists of 9.8k videos, with an average duration of 30 seconds. The dataset includes a total of 157 action categories that are densely labeled in the videos. Charades-STA [2], an extension of Charades, introduces natural language descriptions as annotations. Charades-STA consists of 6.6k videos, with the same average duration as Charades. (3) **ANet and ANet-Caption.** ActivityNet (ANet) [1] contains 200 action categories and ∼20,000 videos. Additionally, ActivityNet-Caption (ANet-Caption) [5] is built upon ANet, including descriptions of events and their corresponding timestamps. ANet-Caption also comprises ∼20,000 videos, having the same average duration of 118 seconds as ANet.

However, **the proportion of overlapping videos between the TAD and MR tasks varies across datasets**, as shown in Table A. In Ego4D, there is a minimal overlap of only ∼2%. Conversely, ANet exhibits a substantial overlap, with more than 99% of the videos being shared. In the Charades, all of the videos from Charades-STA are included, while Charades-STA covers 67% of the videos in Charades. Since each pair of datasets shares the same video source, the duration of the videos in TAD is approximately identical to that in MR. In addition, there are notable differences in the number of instances and average instance duration. In Ego4D, despite having similar average instance

Table A: Statistics of three paired datasets that cover both TAD and MR tasks. #clip denotes the number of video clips, $\#\text{clip}_{\text{diff}}$ presents the number of non-overlapping videos, and $\#\text{clip}_{\text{overlap}}$ refers to the number of overlapping videos. Avg_{dura}, $\text{Avg}_{\text{ins\_dura}}$, and Avg_{ins} indicate the average video duration, the average instance duration, and the average number of instances per video, respectively. #query represents the number of queries in MR while #cls denotes the number of action categories in TAD. The number underlined in ActivityNet indicates that the validation set in TAD only has one version, while ActivityNet-Caption has two versions, val_1 and val_2.

	Temporal Action Detection							Moment Retrieval					
	#clip	$\#\text{clip}_{\text{diff}}$	#cls	Avg_{dura}(s)	Avg_{ins}	$\text{Avg}_{\text{ins\_dura}}$(s)	$\#\text{clip}_{\text{overlap}}$	#clip	$\#\text{clip}_{\text{diff}}$	#query	Avg_{dura}(s)	Avg_{ins}	$\text{Avg}_{\text{ins\_dura}}$(s)
Ego4D													
train	1.5k	1.5k					17	1k	1k	11.3k			
val	0.5k	0.5k					7	0.3k	0.3k	3.9k			
total	2k	2k	110	472	9.2	42.8	**24**	1.3k	1.3k	15.2k	494	10.2	10.6
Charades													
train	8k	2.6k					5.3k	5.3k	0	12.4k			
val	1.8k	0.5k					1.3k	1.3k	0	3.7k			
total	9.8k	3.1k	157	30	6.8	12.9	**6.6k**	6.6k	0	16.1k	30	2.4	8.3
ActivityNet													
train	1w	15					9972	1w	37	3.7w			
val_1	0.5w	9					4917	0.5w	0	1.7w			
val_2		41					4885	0.5w	0	1.7w			
total	2w	65	200	118	1.5	49	**19774**	2w	37	7.1w	118	3.5	35.5

numbers, there is approximately a 300% difference in average instance duration. In Charades, TAD has an average instance number 2.8 times that of MR, and its average instance duration is 50% longer than that of MR. ANet presents a different situation, where MR has a higher average instance number compared to TAD, but the average instance duration is shorter in MR.

Annotation visualizations. Figure A displays six additional annotation visualizations that illustrate the relationship between TAD and MR.

B Details About Implementation

Network architecture details. The proposed unified network for moment detection is described in Sec. 4. For the ablation study on the vision encoder (in Sec. C), we conduct an experiment by replacing the ConvNext [6] blocks with Transformer units, following the same pipeline. In both classification head and regression head, the dimension of the first two convolutional layers is set to 512. In addition, the query transformation branch of the regression head includes three fully connected layers with a dimension of 512. Other parameters, such as the regression range for feature pyramid, center sampling strategy, and EMA model [4], are set according to [8].

Training details. The training of three paired datasets has a consistent AdamW optimizer, weight decay of 0.05, and loss balance weights ($\lambda_{tad} = 3$, $\lambda_{mr} = 1$). However, there are slight differences in other settings. For Ego4D, the model is trained for 15 epochs with a warm-up period of 5 epochs. The batch size is set to 2 and the learning rate is set to 1e-4. In the case of Charades, the model undergoes 20 training epochs, with 10 epochs as a warm-up, a batch size of 2, and a learning rate of 1e-4. For ANet, it is trained for 15 epochs with an initial learning rate of 1e-3, a batch size of 8, and a warm-up period of 5 epochs.

Table B: Effect of different network components evaluated on the TAD task of Ego4D validation set under individual training. The TAD task is the Moment Query task in Ego4D. The $\#D_{reg}$ means the dimensions in the query transformation branch of the regression head and the $scale_{cls}$ indicates the scale operation in the classification head.

Component			TAD	
Encoder	$\#D_{reg}$	$scale_{cls}$	mAP	R1@50
ConvNext	128	Y	21.86	38.79
	256		22.19	40.62
	512		**22.61**	**41.18**
	512	N	22.05	38.58
Transformer	512	Y	21.40	38.90

Table C: Effect of the three solutions for TAD evaluated on the validation set in ANet and val_2 split in ANet-Caption. Here, the methods referred to as "TAD+MR" utilize task fusion learning with random task sampling. The methods "TAD" and "MR" are dedicated models for each task. Solutions $2^{\dagger}$ and $3^{\dagger}$ utilize score fusion strategy.

Method	solution	TAD		MR	
		mAP	mAP@50	R5@50	R5@70
MR	-	-	-	77.28	53.86
TAD	1	37.78	57.48	-	-
TAD+MR		37.98	57.48	79.90	53.71
TAD	$2^{\dagger}$	38.60	58.31	-	-
TAD+MR		**39.82**	**60.04**	**80.83**	**57.21**
TAD	$3^{\dagger}$	38.16	57.42	-	-
TAD+MR		39.79	59.83	80.03	56.89

C Additional Ablation Experiments

In this section, we explore the model settings of the vision encoder and decoder based on TAD metrics of Ego4D, as presented in Table B. Additionally, we investigate three solutions for ANet (Table C) and analyse the effect of task fusion learning on Charades (Table D) and ANet (Table E).

The choice of encoder and decoder. As shown in Table B, using ConvNext as the main block of the vision encoder yields better performance (22.61% vs 21.40% in mAP) compared to Transformer. Additionally, we conduct an exploration of the query transformation branch in the regression head and the learnable scaling operation in the classification head. The comparison indicates that using a Multi-Layer Perceptron (MLP) with a channel size of 512 leads to superior semantic translation, resulting in a higher mAP. Furthermore, the integration of learnable scaling operations positively affects performance.

Three solutions for ANet. In the context of score fusion strategy in ANet, we explore three solutions for TAD, as outlined in Sec. 6.2. Based on the comparison presented in Table C, the score fusion strategy, implemented in solutions 2 and 3, enhances the performance of TAD compared to solution 1 which does not use external classification scores. Besides, among the results of task fusion learning, this approach offers an advantage. Specifically, solution 2 which averages the

Table D: Effect of task fusion learning evaluated on Charades and Charades-STA test set.

Method	policy	TAD	MR			
		mAP	R1@50	R1@70	R5@50	R5@70
TAD	dedicated	22.31	-	-	-	-
MR		-	60.19	41.02	91.61	65.86
MR→TAD	pretrain	22.61	-	-	-	-
TAD→MR		-	62.66	43.15	90.86	64.49
TAD+MR	Random	23.82	63.41	42.42	**92.34**	**67.74**
TAD+MR	Sync.	**24.06**	63.90	42.22	92.12	67.23
TAD+MR	Alt.	21.36	**64.89**	**43.76**	92.02	66.48

Table E: Effect of task fusion learning evaluated on validation set of ANet for TAD and val_2 set of ANet-Caption for MR.

Method	policy	TAD		MR	
		mAP	mAP@50	R5@50	R5@70
TAD	dedicated	38.60	58.31	-	-
MR		-	-	77.28	53.86
MR→TAD	pretrain	39.11	59.46	-	-
TAD→MR		-	-	80.68	55.98
TAD+MR	Random	39.82	60.04	**80.83**	**57.21**
TAD+MR	Sync.	**39.83**	**60.29**	80.54	57.04
TAD+MR	Alt.	39.66	59.51	**80.83**	57.08

text embeddings of entire categories, shows slightly better performance (mAP and mAP@50 in TAD, and R5@50 in MR) compared to the manual prompt for action proposals in solution 3.

Task fusion learning on Charades and ANet. We examine the impact of task fusion learning and different sampling methods on Charades (Table D) and ANet (Table E). In terms of pre-training, both Charades and ANet benefit from this approach, except for the R5@50 and R5@70 in the MR task of Charades. Regarding the effect of co-training on these two datasets, our results align with Ego4D's findings in Sec. 6.3. The synchronized task sampling method (referred to as "Sync.") yields significant improvements in the co-training results for both TAD and MR tasks. Specifically, Charades shows an increase of +1.75% mAP in TAD and +3.71% R1@50 in MR, whole ANet sees an increase of +1.23% mAP in TAD and +3.26% R5@50 in MR. Additionally, we explore the effect of the alternating task sampling (referred to as "Alt."). This sampling leads to significant enhancements in MR for both datasets (+4.70% R1@50 of Charades, +3.55% R5@50 of ANet).

D Qualitative Results

Qualitative results. This section showcases the qualitative results obtained from videos that encompass both TAD and MR tasks, as depicted in Figure B. In our work, we convert the predefined action categories of TAD into natural language queries and treat these queries the same way as queries from MR. The qualitative results demonstrate the capability of unified moment detection using natural language descriptions from both TAD and MR.

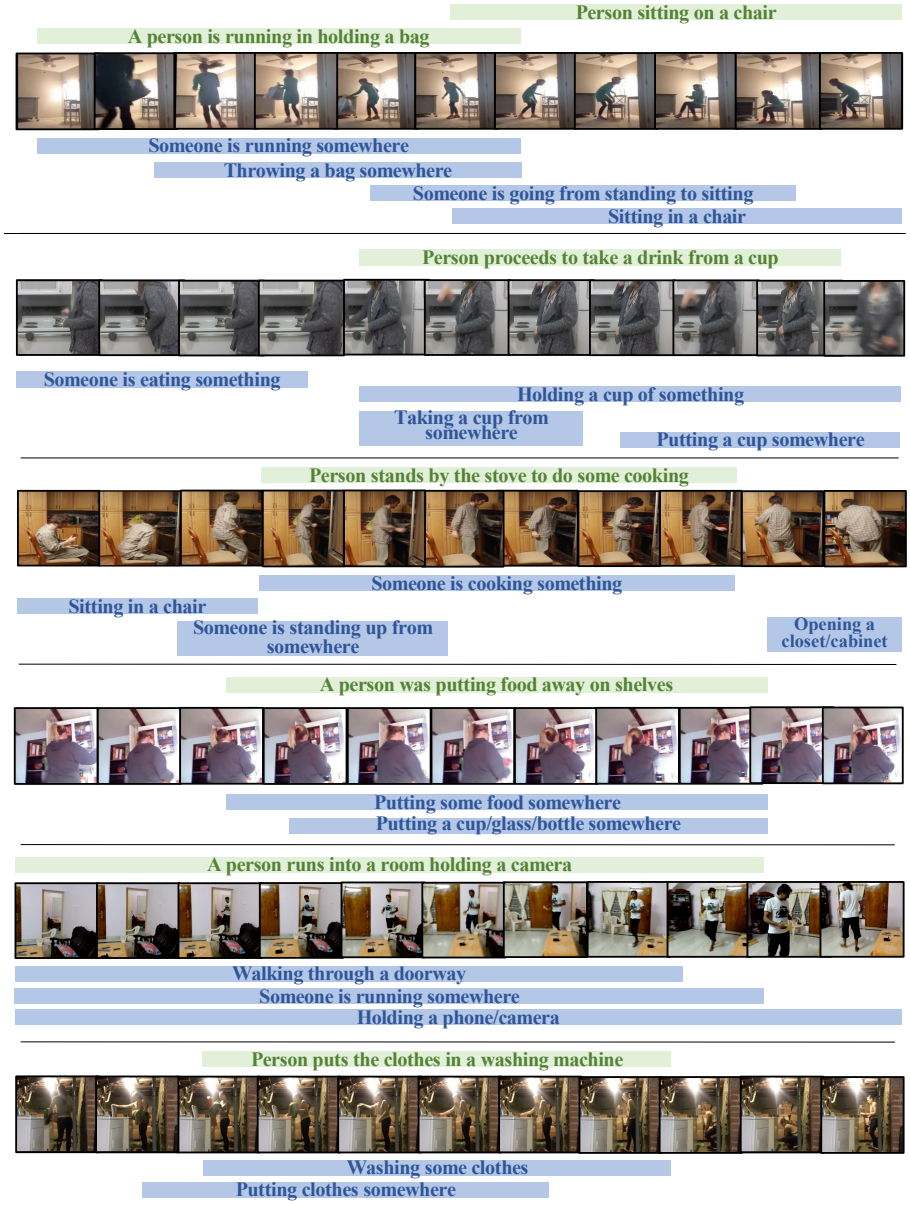
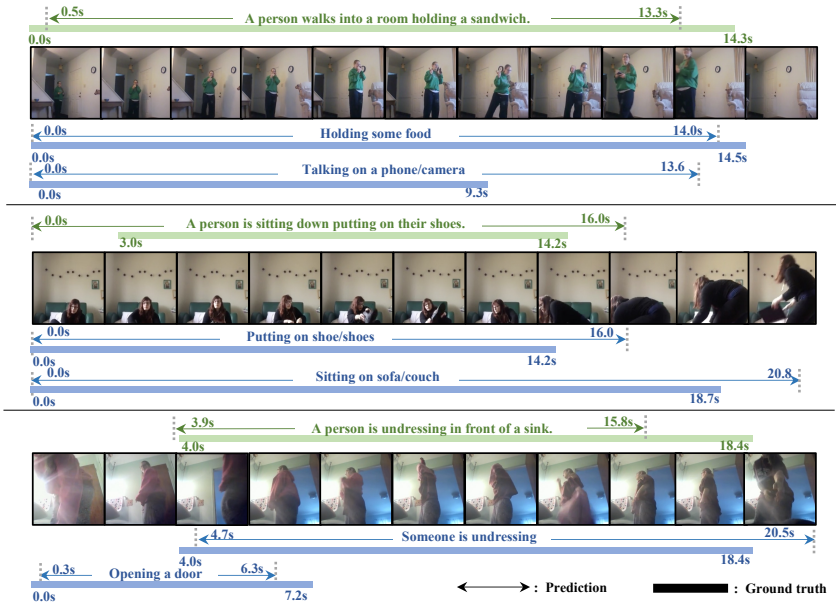
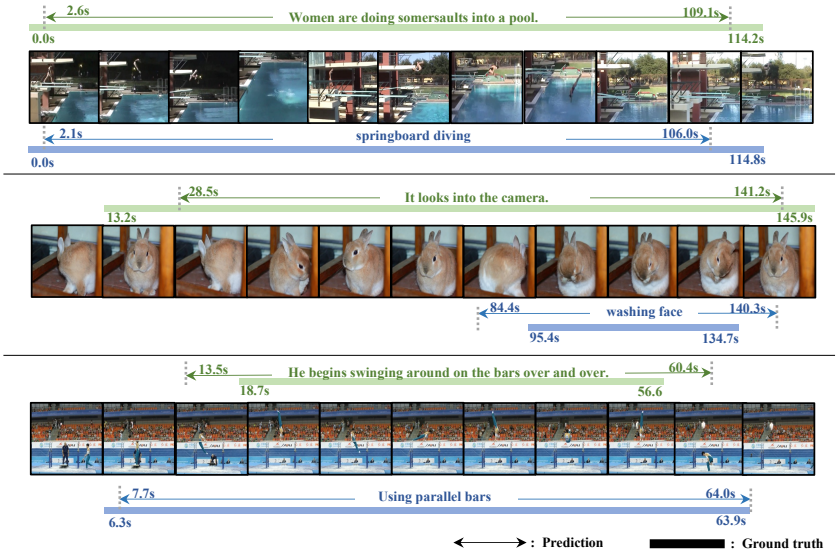


Fig. A: Visualization of annotations from Charades and Charades-STA. The events in green presented above the videos are natural language descriptions from MR, while the actions in blue displayed below the videos belong to predefined categories of TAD.



(a) Qualitative results on Charades and Charades-STA.



(b) Qualitative results on ANet and ANet-Caption.

Fig. B: Qualitative results. The qualitative results provide evidence of the effectiveness of accurate boundary regression and moment detection using natural language descriptions from both MR and TAD. The queries in green presented above the videos are natural language descriptions from MR, while the queries in blue displayed below the videos belong to predefined categories of TAD.

References

1. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the ieee conference on computer vision and pattern recognition. pp. 961–970 (2015)
2. Gao, J., Sun, C., Yang, Z., Nevatia, R.: Tall: Temporal activity localization via language query. In: Proceedings of the IEEE international conference on computer vision. pp. 5267–5275 (2017)
3. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18995–19012 (2022)
4. Huang, G., Li, Y., Pleiss, G., Liu, Z., Hopcroft, J.E., Weinberger, K.Q.: Snapshot ensembles: Train 1, get m for free. arXiv preprint arXiv:1704.00109 (2017)
5. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Carlos Niebles, J.: Dense-captioning events in videos. In: Proceedings of the IEEE international conference on computer vision. pp. 706–715 (2017)
6. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
7. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. pp. 510–526. Springer (2016)
8. Zhang, C.L., Wu, J., Li, Y.: Actionformer: Localizing moments of actions with transformers. In: European Conference on Computer Vision. pp. 492–510. Springer (2022)