Supplementary Material
ProTéGé: Untrimmed Pretraining for Video Temporal Grounding by Video Temporal Grounding

Lan Wang[⋆‡] Gaurav Mittal[⋆†] Sandra Sajeev[†] Ye Yu[†] Matthew Hall[†]
Vishnu Naresh Boddeti[‡] Mei Chen[†]
[†]Microsoft [‡]Michigan State University
{gaurav.mittal, yu.ye, mathall, ssajeev, mei.chen}@microsoft.com
{wanglan3,vishnu}@msu.edu

Below we provide additional qualitative and quantitative analysis for ProTéGé that we could not include in the main paper due to space constraints but was ready at the time of submission. Sec 1 provides a further discussion on hyperparameter selection related to the maximum duration of text query Q, the maximum number of video clips M, and batch size B used during untrimmed pretraining. Next, Sec 2 provides visual analysis for ProTéGé both for pretraining and downstream Video Temporal Grounding (VTG), and finally, Sec 3 provides additional implementation details.

1. Additional Discussion

Maximum duration of text query Q. Since our method is not limited to a single subtitle as a text query and uses *aggregated subtitles* by concatenating them, we discuss the effect of maximum query length in terms of duration in Table 1. We can observe that a maximum query duration of 50 seconds gives the best performance. We believe shorter durations limit the generalization of the method to downstream tasks with diverse query sizes. Meanwhile, a duration of 100 seconds is sub-optimal because by then, the query has too much information by having as many as 25 subtitles. This reduces its usefulness to precisely localize and be associated with a particular video segment.

Table 1. Effect of maximum duration of text query Q.

Max duration of Q (s)	R@0.5	R@0.7
10	52.59	27.48
20	50.19	26.83
50	53.26	30.38
100	48.27	27.73

Maximum number of video clips M. The video duration, in terms of the number of clips M, can play a significant role in the performance. As shown in Table 2, using at most $M = 15$ or $M = 30$ video clips significantly reduces R@0.5 by 5.28% and 7.07% respectively compared to using $M = 60$. This suggests that using long untrimmed videos makes the pretrained features more favorable for downstream VTG tasks. We also find that increasing M from 60 to 120 does not provide a significant improvement. Larger M results in more fine-grained proposals which are quadratically more in number. We believe that this increases task complexity making the training more challenging while also requiring longer training time as well as higher GPU memory. So for compute efficiency, we use $M = 60$ in our experiments.

Table 2. Effect of maximum number of video clips M.

Max video clips M	R@0.5	R@0.7
15	46.19	25.14
30	47.98	25.59
60	53.26	30.38
120	53.24	30.60

Table 3. Effect of batch size B during pretraining.

Batch size B	R@0.5	R@0.7
512	50.67	29.11
1024	52.51	30.86
2048	53.26	30.38
4096	51.74	28.69

[⋆]Authors with equal contribution.
This work was done as Lan Wang's internship project at Microsoft.

Batch size B during pretraining. The batch size B during untrimmed pretraining of ProTéGé decides the number of negative samples in $\mathcal{L}_{inter}$ and influences model training. Table 3 shows the results for using different batch sizes for pretraining. We find that using $B = 2048$ gives the overall best performance and having smaller or larger batch size leads to worse performance. We believe that having a smaller batch size can cause the model to have an insufficient number of negative samples while having a larger batch size could lead to a high number of false negatives. Both of these scenarios can impede model training [8].

Effectiveness of VT-SGM. We directly incorporate our proposed grounding module (VT-SGM) into pretraining to leverage untrimmed videos for VTG. While downstream VTG methods like 2D-TAN [9] inspire VT-SGM, our module's design is tailored to perform VTG-based untrimmed pretraining. To illustrate the effectiveness of VT-SGM, we replace VT-SGM in ProTéGé with the original 2D-TAN and compare the performance on Charades-STA and TACoS on fully-supervised VTG as the downstream task in Table 4. We can observe that using VT-SGM leads to 2.7%/4.0% higher R@0.5/R@0.7 on Charades and 1.8%/1.8% higher R@0.3/R@0.5 on TACoS than original 2D-TAN module. This further validates the benefit of the novel design of our VT-SGM module.

Table 4. Comparison of using VT-SGM vs. original 2D-TAN module in VTG pretraining. VT-SGM significantly outperforms original 2D-TAN module on both Charades-STA and TACoS on downstream task for Fully-supervised VTG.

Grounding Module	Charades-STA		TACoS	
	R@0.5	R@0.7	R@0.3	R@0.5
2D-TAN [9]	50.53	26.40	41.88	29.36
VT-SGM	**53.26**	**30.38**	**43.63**	**31.39**

Evaluation on more VTG datasets. Table 6 and 5 show results on TACoS [5] and QVHighlights [2] using ProTéGé with Moment-DETR and 2D-TAN as downstream methods for fully-supervised VTG. We can observe that untrimmed pretraining (Row 3) leads to 2.1%/4.2% better R@0.5/R@0.7 on QVHighlights and 1.8%/1.2% better R@0.3/R0.5 on TACoS vs. our baseline of ProTéGé without untrimmed pretraining (Row 2) using the same extra video data, which empirically validates our untrimmed pretraining algorithm for VTG. Moreover, compared with the baselines, ProTéGé also shows superior performance.

Evaluation on VidSitu (Video Event Relation Understanding). To demonstrate the effectiveness of ProTéGé on understanding movie data, we further evaluate ProTéGé on VidSitu [6] which is a large-scale dataset containing diverse videos from movies depicting complex situations. Specifically, we choose the event relation classification as our downstream task and Vid TxEnc [6] as the downstream method. Table 7 shows macro-average Accuracy (Macro-Acc) on the validation set. We can observe that ProTéGé (Row 3) exceeds both Vid TxEnc (Row 1) and our baseline of ProTéGé without untrimmed pretraining (Row 2), further validating its generalization ability on movie data and complex situation understanding tasks.

Table 5. Evaluation on QVhighlight in Fully-supervised VTG downstream setting. ProTéGé (Row 3) significantly outperforms both Moment-DETR (Row 1) and the baseline without untrimmed pretraining (Row 2).

Method	R@0.5	R@0.7
Moment-DETR [2]	53.94	34.84
ProTéGé w/o Untrimmed	53.53	31.91
ProTéGé	**55.56**	**36.11**

Table 6. Evaluation on TACoS in Fully-supervised VTG downstream setting. ProTéGé (Row 3) significantly outperforms both LocVTP (Row 1) and the baseline without untrimmed pretraining (Row 2).

Method	R@0.3	R@0.5
LocVTP [1]	41.6	28.9
ProTéGé w/o Untrimmed	41.76	30.01
ProTéGé	**43.63**	**31.19**

Table 7. Evaluation on VidSitu. ProTéGé (Row 3) significantly outperforms both Vid TxEnc (Row 1) and the baseline without untrimmed pretraining (Row 2).

Method	Macro-Acc
Vid TxEnc [6]	34.54
ProTéGé w/o Untrimmed	41.81
ProTéGé	**45.47**

2. Additional Qualitative Analysis

Fig 1 and Fig 2 show the similarity score proposal grid on videos from the Charades-STA and ActivityNet-Captions respectively for ProTéGé and a baseline setup doing pretraining on trimmed videos. Both setups have never seen the videos during pretraining. We feed the videos and text queries through the video and text query encoders respectively and using the output features, obtain the cosine similarity scores for the 2D proposal grid without doing any

finetuning on the videos. The cyan dot in the grid in the figures denotes the location of the ground truth proposal for the corresponding query. We can observe that ProTéGé, having been trained on untrimmed videos, can clearly learn to exhibit higher video-text similarity close to the ground truth and lower similarity farther away from the ground truth. But when trained on trimmed videos, there is no visible difference in the similarity scores across proposals. Moreover, the range of similarity scores is also significantly larger for ProTéGé. This shows that our method, pretrained on untrimmed videos, can develop a more fine-grained understanding of the video, leading to more discriminative intra-video features and allowing for more distinguishable video-text similarity across different regions within a video.

Fig 3 and Fig 4 further compare the fully supervised VTG performance on videos from the Charades-STA and ActivityNet-Captions respectively by visualizing the localization results. We use 2D-TAN [9] as the downstream method and compare ProTéGé with a baseline setup using features from backbone pretrained on trimmed videos. For both datasets, ProTéGé features can ground the text query in the video significantly more precisely while features pretrained on trimmed videos exhibit a large deviation from the ground truth when grounding the query in the video. As shown in Fig 3c, Fig 3d, Fig 4b, and Fig 4c, when the background inside the ground truth location is visually very similar to the outside background, the baseline makes large errors in correctly grounding the query in the video. On the other hand, ProTéGé, due to its ability to learn highly discriminative features within a video via pretraining on untrimmed videos, is able to perform significantly better and provide accurate query localization.

3. Additional Implementation Details

Expanding on the implementation details in the main text, we conduct our experiments using Swin-T [4] pretrained on Kinetics-400 and Swin-B [4] pretrained on Kinetics-600 as the frozen trimmed video encoders, f^{vf}. From Table 1 of the main text, we can see significant improvements on both backbones using ProTéGé that highlights the usefulness of our method across backbones of different sizes. We use Hugging Face's [7] implementation of RoBERTa-base [3] for the frozen text encoder f^{qf}. For downstream VTG tasks, we only use the visual features from our video encoder f^{v} to have a fair comparison with existing methods. For videos longer than 128s, we use a non-overlapping sliding window of 128s for feature extraction. We resize video frames to 224×224 to feed to the video encoder. Before tokenizing the text, we clean it by lower-casing the text, de-accenting the characters, and removing unicode characters, punctuation, and stop words.

References

[1] Meng Cao, Tianyu Yang, Junwu Weng, Can Zhang, Jue Wang, and Yuexian Zou. Locvtp: Video-text pre-training for temporal localization. In *European Conference on Computer Vision*, pages 38–56. Springer, 2022. 2

[2] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34:11846–11858, 2021. 2

[3] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 3

[4] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022. 3

[5] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36, 2013. 2

[6] Arka Sadhu, Tanmay Gupta, Mark Yatskar, Ram Nevatia, and Aniruddha Kembhavi. Visual semantic role labeling for video understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 2

[7] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, Oct. 2020. Association for Computational Linguistics. 3

[8] Jun Xia, Lirong Wu, Ge Wang, Jintao Chen, and Stan Z. Li. ProGCL: Rethinking hard negative mining in graph contrastive learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 24332–24346. PMLR, 17–23 Jul 2022. 2

[9] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. 2, 3

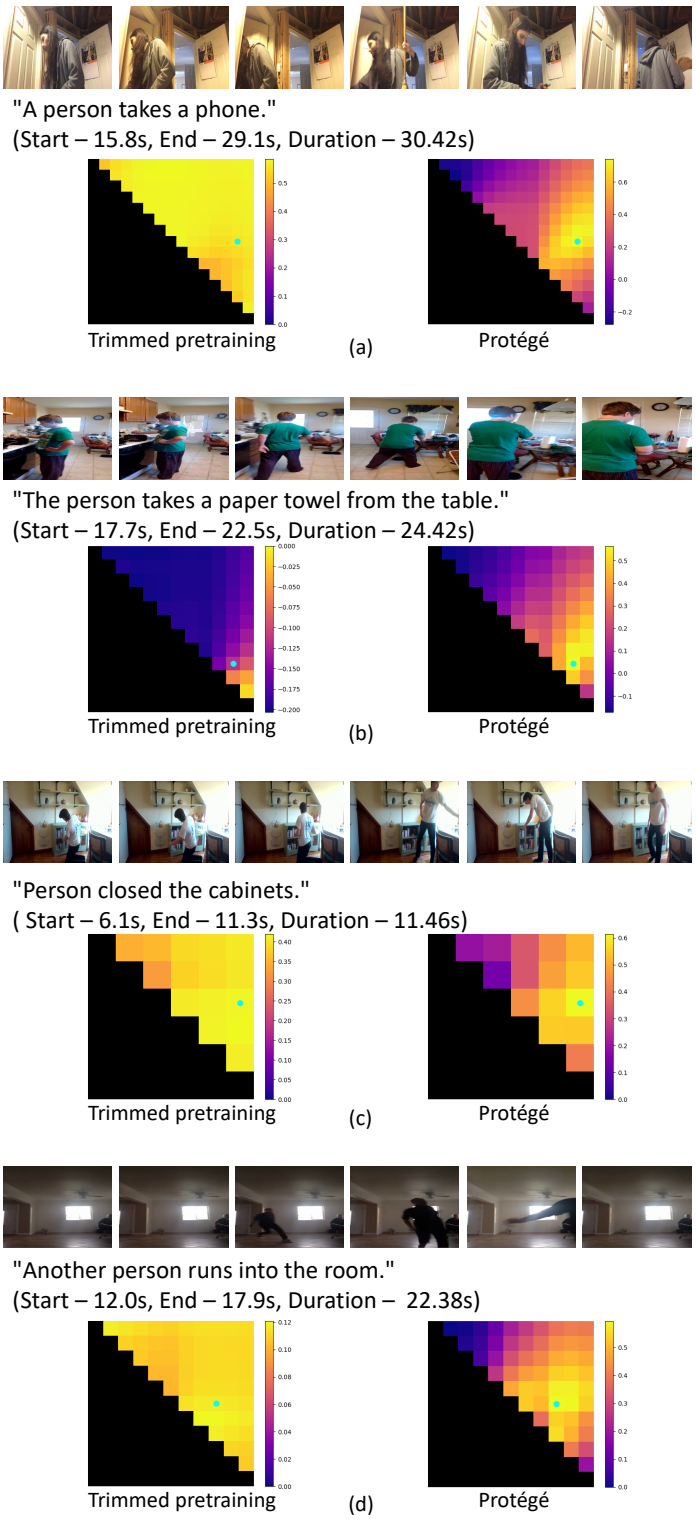
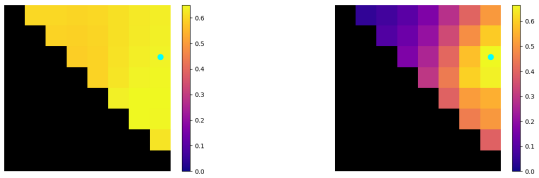


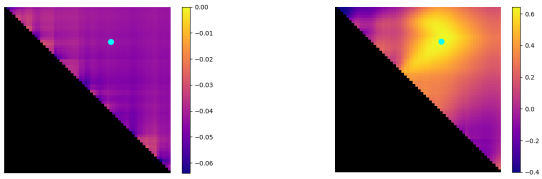
Figure 1. Visualization of 2D proposal grid cosine similarity scores on unseen Charades-STA videos. For each example, the first row shows the video frames and the second row provides the text query along with start-end timestamp and video duration. The third row compares the 2D proposal grid cosine similarity scores of a baseline pretrained on trimmed videos (left) with ProTéGé pretrained on untrimmed videos (right). ProTéGé features show higher variation in cosine similarity with larger similarity closer to the ground truth (cyan dot) due to ProTéGé's ability to learn discriminative features within a video. The grid size varies based on the length of the untrimmed video.

"Afterwards the camera pans around to a small group of men making plaster." (Start – 5.64s, End – 16.83s, Duration – 17.09s)



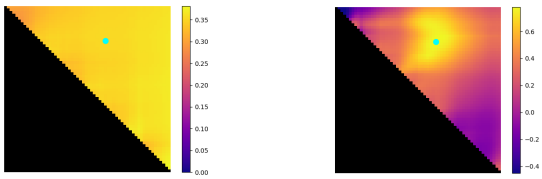
Trimmed pretraining (a) Protégé

"They twirl around on the dance floor in circles."
(Start – 38.71s, End – 119.83s, Duration – 184.36s)



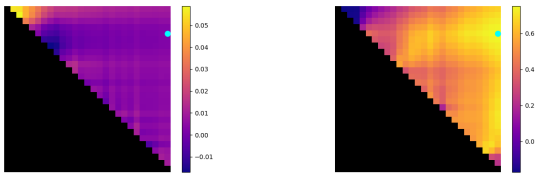
Trimmed pretraining (b) Protégé

"The man stopped walking and talked to the man lying on the tube."
(Start – 30.84s, End – 89.57s, Duration – 146.84s)



Trimmed pretraining (c) Protégé

"They play their instruments together."
(Start – 5.72s, End – 57.24s, Duration – 57.24)



Trimmed pretraining (d) Protégé

Figure 2. Visualization of 2D proposal grid cosine similarity scores on unseen ActivityNet-Captions videos. For each example, the first row shows the video frames and the second row provides the text query along with start-end timestamp and video duration. The third row compares the 2D proposal grid cosine similarity scores of a baseline pretrained on trimmed videos (left) with ProTéGé pretrained on untrimmed videos (right). ProTéGé features show higher variation in cosine similarity with larger similarity closer to the ground truth (cyan dot) due to ProTéGé's ability to learn discriminative features within a video. The grid size varies as per the length of the untrimmed video.

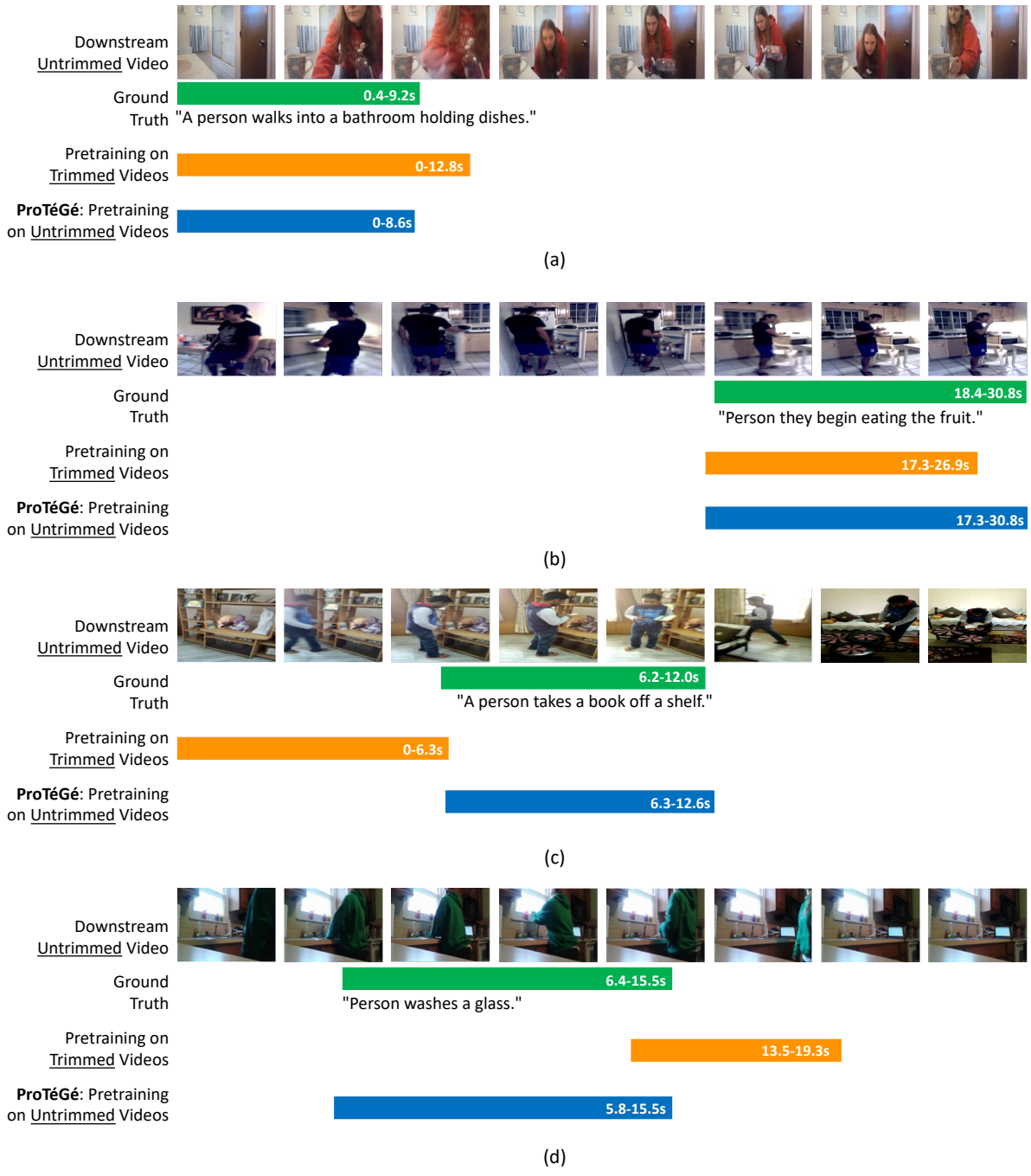


Figure 3. Visualization of fully-supervised video temporal grounding results on Charades-STA dataset using 2D-TAN as the downstream method. For each example, the first row shows the frames of the untrimmed video, the second row shows the ground truth location of the query in the untrimmed video in green, the third row shows grounding prediction in orange from a baseline pretrained on trimmed videos, and the fourth (final) row shows grounding prediction in blue from ProTéGé pretrained on untrimmed videos. We can observe that ProTéGé shows more accurate grounding predictions for all examples as it is pretrained on untrimmed videos which allows ProTéGé to develop a more fine-grained understanding of the video and learn more discriminative features within a video.

Figure 4. Visualization of fully-supervised video temporal grounding results on ActivityNet-Captions dataset using 2D-TAN as the downstream method. For each example, the first row shows the frames of the untrimmed video, the second row shows the ground truth location of the query in the untrimmed video in green, the third row shows grounding prediction in orange from a baseline pretrained on trimmed videos, and the fourth (final) row shows grounding prediction in blue from ProTéGé pretrained on untrimmed videos. We can observe that ProTéGé shows more accurate grounding predictions for all examples as it is pretrained on untrimmed videos which allows ProTéGé to develop a more fine-grained understanding of the video and learn more discriminative features within a video.