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

Abstract

Video temporal grounding (VTG) is the task of localizing a given natural language text query in an arbitrarily long untrimmed video. While the task involves untrimmed videos, all existing VTG methods leverage features from video backbones pretrained on trimmed videos. This is largely due to the lack of large-scale well-annotated VTG dataset to perform pretraining. As a result, the pretrained features lack a notion of temporal boundaries leading to the video-text alignment being less distinguishable between correct and incorrect locations. We present ProTéGé as the first method to perform VTG-based untrimmed pretraining to bridge the gap between trimmed pretrained backbones and downstream VTG tasks. ProTéGé reconfigures the HowTo100M dataset, with noisily correlated video-text pairs, into a VTG dataset and introduces a novel Video-Text Similarity-based Grounding Module and a pretraining objective to make pretraining robust to noise in HowTo100M. Extensive experiments on multiple datasets across downstream tasks with all variations of supervision validate that pretrained features from ProTéGé can significantly outperform features from trimmed pretrained backbones on VTG.

1. Introduction

Video temporal grounding (VTG) is the video-language multimodal task of localizing which part of an arbitrarily long untrimmed video can be best associated with a given natural language text query. VTG has a wide range of applications, such as information retrieval and robotics. Figure 1 shows a sample video-text pair for the VTG task and illustrates the primary challenge in grounding an unconstrained natural language text query in a long untrimmed video, namely, the need for a fine-grained understanding of the spatio-temporal dynamics in the video.

*Authors with equal contribution.
This work was done as Lan Wang’s internship project at Microsoft.

Figure 1. (a) Comparison between video-text trimmed and untrimmed pretraining on grounding text Q1 and Q2 in an untrimmed video. Untrimmed video-text pretraining shows stronger grounding capability. (b) and (c) show box plots of cosine similarity after joint video-text pretraining on trimmed and untrimmed videos, respectively between video features aligning with text (blue) and not aligning with text (red). We observe that compared to using trimmed videos (b), cosine similarities of video features aligning with text are higher and farther apart from that of video features not aligning with text when using untrimmed videos (c), thus illustrating the impact of untrimmed pretraining.

While there are multiple approaches for VTG, all existing methods, to the best of our knowledge, rely on video backbones pretrained on trimmed videos (such as Kinetics [15]) to obtain the visual features as part of their respective approaches. Such a design choice introduces a disconnect between the downstream VTG task on untrimmed videos and the trimmed videos used for pretraining the model from which video features are derived. For example, Fig 1 shows that the grounding predictions (in orange), when using a backbone jointly pretrained on trimmed videos and text, do not match adequately with the ground truth. Due to pretraining on trimmed videos, the video backbone is insensitive to temporal boundaries since the training objective is to associate an entire trimmed video to a label/text
query [43, 44]. The backbone, therefore, does not have an explicit ability to localize, i.e., associate the given query to only the most relevant part of the long untrimmed video. As a result, the cosine similarity between video features aligning and not aligning with the text query are indistinguishable, as shown in Fig 1b.

Inspired by the advantage shown in other tasks where pretraining and downstream setup match [3, 7, 23, 26], we hypothesize that formulating the pretraining itself as a VTG task on untrimmed videos can improve downstream grounding performance. The untrimmed pretraining will equip the model with a more accurate and fine-grained understanding of temporal boundaries within a given untrimmed video (as evidenced by the more precise predictions in blue in Fig. 1). We introduce ProTeGe, Untrimmed Pretraining for Video Temporal Grounding by Video Temporal Grounding. ProTeGe is the first approach to formulate pretraining as a VTG task to bridge the gap between video backbones pretrained on trimmed videos and downstream VTG tasks working with untrimmed videos.

A critical challenge impeding this untrimmed pretraining is the scarcity of large-scale well-annotated video grounding datasets. There are, however, datasets such as HowTo100M [30] and YouTube-8M [1], with over a million untrimmed videos and corresponding subtitled text generated via automated speech-to-text APIs. One can potentially employ them for untrimmed pretraining as a VTG task. However, as noted by prior methods [12, 29, 41], since the text is derived from subtitles, the video regions are only noisily-correlated with the subtitled text, rendering the utility of these video-text pairs for grounding a non-trivial task.

To overcome the aforementioned challenges in leveraging large-scale untrimmed video datasets, we first propose a novel approach to transform them into VTG datasets. Then we introduce a novel video-text similarity grounding module along with an optimization objective that allows the pretraining to be robust to the noisy video-text correlations present in these datasets.

In this work, we use ProTeGe with HowTo100M in particular. To transform HowTo100M into a VTG dataset, ProTeGe introduces aggregated subtitles to concatenate one or more subtitles to form the text query and randomly samples an untrimmed video segment around the query. Aggregated subtitles allow ProTeGe to incorporate arbitrarily long text queries larger than the average 4s duration of a single subtitle. This way, we can synthesize millions of video-text grounding pairs for VTG pretraining. Using these pairs, ProTeGe performs pretraining with our novel Video-Text Similarity-based Grounding Module (VT-SGM). VT-SGM creates a 2D-proposal grid by computing the cosine similarity between the text query and the different temporal regions of the untrimmed video. It then learns to maximize the similarity between the query and the part that is most relevant to it. This is achieved via our novel pretraining objective that incorporates a distance-based pretraining objective that uses the noisy ground truth and a combination of inter-video and intra-video alignment losses. This allows the objective to balance the training via the noisy ground truth and multimodal video-text representation learning. We show that ProTeGe is very effective for VTG as a downstream task. It significantly outperforms backbones pretrained on trimmed videos on standard datasets across all variations of supervision. We summarize our contributions as,

1. We propose ProTeGe, the first pretraining method formulated as a video temporal grounding task to bridge the gap between pretraining and downstream video temporal grounding tasks in untrimmed videos.

2. We propose a novel algorithm including aggregated subtitles, a Video-Text Similarity-based Grounding Module, and a pretraining objective to leverage large-scale untrimmed video dataset HowTo100M with noisy video-text pairs.

3. Extensive experiments on standard datasets across multiple downstream tasks with different levels of supervision validate that our approach significantly improves the performance across all benchmarks.

2. Related Work

Video Temporal Grounding. First proposed in [2, 8], VTG aims to retrieve the temporal moment for the given sentence [46, 50]. Most works, including [2, 8, 20, 21, 32, 48, 51], are developed in a supervised setting, which utilizes the start and end timestamp as supervision. Some methods use reinforcement learning to address the problem [11, 13, 39]. Those methods formulate VTG as a sequential decision-making problem. Moreover, weakly supervised VTG is also proposed which only utilizes the sentence and video without any localization annotations, achieving competitive results [9, 19, 28, 31, 37]. Other works like PSVL [33] and DSCNet [22] proposed VTG in a zero-shot or unsupervised manner without seeing supervised information.

Video-Language Pre-training. Benefiting from multimodality and large-scale datasets, significant works have studied pre-training tasks using both video and language [10, 14, 18, 27, 29, 34, 35, 40, 41, 47, 55]. Videobert [36] extracted a set of cooking videos from YouTube, and adapts the BERT model, learning a joint visual-linguistic representation. Clipbert [17] utilized image-text datasets for pre-training, and then fine-tune on the video-text downstream task. With the introduction of HowTo100M [30], a video dataset with more than 130M video clips and corresponding transcription, many works leverage the clip-caption pairs to learn joint text-video embedding. MIL-NCE [29] proposed a multiple-instance learning objective to deal with
the misaligned narration descriptions from Howto100M. Multimodal Pre-training (MMP) [14] extended Howto100M to a multilingual dataset to mitigate the performance gap for non-English data. VLM [40] used a single BERT encoder to realize a task-agnostic pretraining, which can accept single/multiple modalities for different downstream tasks. Videoclip [41] utilized temporally overlapped pairs to learn fine-grained associations between video frames and word tokens. TAN [12] introduced a temporal-alignment network that targeted refining alignable text in Howto100M [12]. Most existing pretraining works are focused on solving downstream tasks like text-video retrieval [2], VideoQA [42], video captioning [54], and action recognition [4] under zero-shot or finetune settings.

Pre-training for localization. Several works study pre-training for localization tasks. Lofi [44] proposed to jointly optimize the video encoder from trimmed pretraining and TAL head, resolving the discrepancy problem for downstream tasks. PAL [49] designed a self-supervised pretext task for temporal action localization, achieving unsupervised pretraining. BSP [43] proposed a video synthesis method that used 4 different boundary strategies to generate videos with temporal boundary information to facilitate pretraining. LocVTP [5] combined coarse-grained and fine-grained contrastive loss and utilized a temporally aware contrastive learning to optimize the pretraining for temporal localization tasks. Although specifically designed for localization-related downstream tasks, those methods are still pretrained in a trimmed manner, which is sub-optimal for localization downstream tasks.

3. Method

Let \( V = [v]^T \) be an arbitrarily long untrimmed video where \( v_1, \ldots, v_T \) denote the sequence of \( T \) frames forming the video. Let \( Q = [q_1]^L \) denote a text query comprising a sequence of \( L \) word tokens, \( q_1, \ldots, q_L \). Since ProTeGé formulates the pretraining as an untrimmed VTG task to reduce the discrepancy between pretraining and downstream VTG, we formulate pretraining as localizing the start-end timestamp tuple \((S, E)\) in the video \( V \) that best matches the textual description in query \( Q \).

3.1. Synthesizing VTG Dataset for Pretraining

Since large-scale untrimmed video datasets with temporally well-annotated captions are unavailable for pretraining as a VTG task, we leverage the video dataset HowTo100M for VTG-based untrimmed pretraining in ProTeGé. HowTo100M comprises over a million untrimmed videos and autogenerated speech-to-text subtitles. The subtitles can potentially serve as text queries \( Q \) in our VTG-based untrimmed pretraining. On top of the challenge of noisy video-text correlation which we address...
in Sec 3.3, another key obstacle in leveraging such datasets is that the duration of a single subtitle text is very small (e.g., 4s on avg. for HowTo100M while downstream VTG datasets can have arbitrarily long captions (e.g., 37s on avg. for ActivityNet-Captions [16]), leading to a discrepancy and sub-optimal performance.

We, therefore, propose to use **aggregated subtitles** where we concatenate one or more captions together to form the query. Specifically, for an untrimmed video \( V \), let \( SB = \{SB_j \}_{j=1}^n \) be the sequence of \( U \) subtitles. For the \( K^{th} \) sample of our synthesized VTG dataset, we form query \( Q_K \) as \( Q_K = \{SB_j \}_{j=m}^n = [q_{m1}, \ldots, q_{mL_{SB_m}}, \ldots, q_n, \ldots, q_{nL_{SB_n}}] \), where \( 1 \leq m \leq n \leq U \) and \( L_{SB_m} \) and \( L_{SB_n} \) are the number of word tokens in subtitles \( SB_m \) and \( SB_n \) respectively. Let \( v_s \) and \( v_e \) be the video frames where \( Q_K \) starts and ends respectively in \( V \). We also randomly sample an arbitrarily long video segment, \( V_K = [v_s^e]_{s=1}^{e} \) with \( v_s \) and \( v_e \) as start and end frames of \( V_K \) respectively such that \( 1 \leq v_s \leq v_s < v_e < v_e \leq T_V \).

We then define the normalized start-end timestamp tuple \( (S_K, E_K) \) where \( Q_K \) temporally grounds in \( V_K \) as,

\[
S_K, E_K = \frac{s - s'}{e' - s'}, \frac{e - e'}{e' - s'}, 0 \leq S_K < Q_K \leq 1 \tag{1}
\]

This enables us to synthesize dataset samples \( \{V_K, Q_K, (S_K, E_K)\} \) for our VTG-based pretraining task (Fig 2a). It is worth noting that contrary to all previous works which leverage datasets like HowTo100M as an independent collection of trimmed video-text pairs where each pair is extracted using the subtitles’ start-end timestamp, our proposed method to synthesize a VTG dataset makes it possible to leverage videos for pretraining as untrimmed with arbitrarily long duration.

### 3.2. Model Overview

To perform VTG during pretraining, ProTéGé consists of a network with a video encoder to extract features for the untrimmed video \( V_K \) and a text query encoder to encode the corresponding text query \( Q_K \). We then use a novel video-text similarity-based grounding module that learns to align the text query with the most relevant part of the video via a novel pretraining objective for VTG. This enables learning features that are better suited for downstream VTG tasks. Fig 2 illustrates an overview of ProTéGé. We omit \( K \) in subsequent text for simplicity.

**Video Encoder.** As shown in Fig. 2b, ProTéGé comprises a video encoder \( f^V(.) \) that takes as input an untrimmed video \( V \) with an arbitrary number of frames \( T_V \). We design \( f^V(.) \) to leverage the feature representations learned by the pretrained video encoder \( f^{\text{tv}}(.) \) trained on a large set of trimmed videos. These features serve as a rich representation of the local temporal neighborhood and help reduce computational overhead by reusing already performed pretraining. Moreover, freezing \( f^{\text{tv}}(.) \) allows \( f^V(.) \) to focus on learning long-term temporal interactions in an untrimmed video, which is missing in \( f^{\text{tv}}(.) \) and is the source of discrepancy between trimmed pretraining and downstream untrimmed VTG. We split arbitrarily long \( V \) into up to \( M \) clips, \( C_V = [C_{Vj}]_{j=1}^M \) where each clip has a fixed \( T_C \) number of frames (padding last clip with its last frame if needed). We first feed each clip to \( f^V(.) \) independently and do an average pooling over the frame dimension to obtain a sequence of clip-level local feature representations \( h_V = [h_{Vj}]_{j=1}^M \) of size \( M \times D \), where \( D \) is the feature dimension. \( h_V \) is then fed to a transformer-based untrimmed video encoder, \( f^{\text{un}}(.) \), to obtain the temporal sequence of feature representations for the untrimmed video as \( z_V = [z_{Vj}]_{j=1}^M \) such that \( z_V = f^V(V) = f^{\text{un}}(f^V(V)) \).

**Query Encoder.** To encode the text query \( Q \), ProTéGé comprises a query encoder \( f^Q(.) \) whose design is symmetric to the video encoder \( f^V(.) \) (Fig 2b). \( f^Q(.) \) comprises a frozen pre-trained text encoder \( f^{\text{tq}}(.) \) which first tokenizes the query text via embedding lookup and then processes it to output a sequence of features \( h_Q \). After this, we feed \( h_Q \) through a randomly initialized transformer-based query encoder \( f^{\text{un}}(.) \) which performs an average pooling over the features for each at the end to obtain the final query feature embedding \( z_Q = f^Q(Q) = \text{AvgPool}(f^{\text{un}}(f^{\text{tq}}(Q))) \).

**Video-Text Similarity-based Grounding Module (VT-SGM).** Similar to downstream VTG tasks, we formulate VTG during pretraining as a task to align text with the correct region in the untrimmed video. This explicitly primes the model to specialize for downstream VTG tasks. Unlike video pretraining methods that leverage trimmed videos, we cannot align the text features with the entire video feature sequence. To tackle this, we propose a novel Video-Text Similarity-based Grounding Module (VT-SGM) that learns to localize the text query in the untrimmed video. The module first generates an upper triangular 2D proposal grid, \( \mathbb{PG} \in \mathbb{R}^{M \times M} \), where each grid cell maps to a video proposal (Fig 2b). Taking output \( z_V = [z_{Vj}]_{j=1}^M \) of video encoder \( f^V(.) \), a cell, \( p_{\text{grid}} \), at row \( s \) and column \( e \) in proposal grid \( \mathbb{PG} \) maps to a video proposal spanning the feature sequence \( [z_{Vs}, \ldots, z_{Ve}] \) where \( 1 \leq s \leq e \leq M \). We obtain the feature for video proposal for each cell, \( z_{\text{vp}, s,e} \), by applying an average pooling over the temporal dimension such that \( z_{\text{vp}, s,e} = \frac{\sum_{i=s}^{e} z_{Vi}}{e-s+1} \forall 1 \leq s \leq e \leq M \). Finally, we compute the grid score \( g_{\text{vp}, s,e} \) of each cell \( p_{\text{grid}} \), as cosine similarity of the text query features \( z_Q \), obtained from the query encoder \( f^Q(.) \), with \( z_{\text{vp}, s,e} \) for every \( 1 \leq s \leq e \leq M \) as,

\[
g(s,e) = \frac{z_{\text{vp}, s,e}^T z_Q}{||z_{\text{vp}, s,e}|| \cdot ||z_Q||} \tag{2}
\]

The similarity score \( g_{\text{vp}, s,e} \) reflects how well the region in the untrimmed video represented by \( z_{\text{vp}, s,e} \) aligns with the
query text feature $z_Q$. We take inspiration for the design of $\text{PG}$ from 2D-TAN [51] but we differ significantly in that unlike 2D-TAN, $\text{PG}$ is a 2D grid instead of a 3D temporal feature grid used by 2D-TAN. We use cosine similarity to obtain the grid scores whereas 2D-TAN applies elementwise multiplication on video-text features followed by a series of convolutional layers to obtain the final grid scores. We find our approach less complex and therefore more robust (Sec 4) in the presence of data from sources like HowTo100M with noisy video-text correlations.

### 3.3. Pretraining Objective for Temporal Grounding

One major challenge in leveraging the synthesized VTG dataset from HowTo100M (Sec 3.1) is that the subtitles may not necessarily align with the video. To enable our model to be robust to these noisy video-text correlations, we design a novel pretraining objective for VTG which augments what the model learns from the noisy ground truth $SE = (S, E)$ with what the model can learn implicitly via multimodal video-text representation learning. Our pretraining objective, therefore, comprises a localization loss, inter-video-text alignment loss, and intra-video-text alignment loss.

#### Localization Loss.

The localization loss, $L_{\text{loc}}$ leverages the fact that although our ground truth $SE = (S, E)$ is noisy, it can still provide some guidance to the model in terms of approximately what part of the untrimmed video can be best described by the text query. Therefore, one component of this loss is the standard cross-entropy loss, $L_{\text{ce}} = \text{CrossEntropyLoss}(F_V, SE)$, where $F_V \in \mathbb{R}^H$ is the sequence of proposal grid scores $g_{se}$ for $1 \leq s \leq e \leq M$ obtained by flattening the 2D proposal grid, $\text{PG}$ and $H$ is the total number of video proposals, $H = \frac{M(M+1)}{2}$ (Fig 2c).

$L_{\text{ce}}$ serves as a hard-localization loss which does not account for the proximity of the localization from the ground truth and equally penalizes all incorrect localizations. $L_{\text{ce}}$ alone is not optimal in the presence of noisy ground truth that we have. We, therefore, propose an additional component of $L_{\text{loc}}$ that penalizes an incorrect prediction proportionate to how far the localization prediction from the ground truth.

To achieve this soft localization, we compute a distance map, $D \in \mathbb{R}^{M \times M}$, such that $D(s, e)$ is defined as the Manhattan distance between the proposal corresponding to $(s, e)$ and the ground truth $SE = (S, E)$, $D(s, e) = |S - s| + |E - e|$. Similar to $L_{\text{ce}}$, we flatten $D$ to a sequence of distance scores $D_V \in \mathbb{R}^H$ and compute the distance-based soft-localization loss, $L_{\text{dist}}$ as the L1 loss,

$$L_{\text{dist}} = \|F_V, 2 \left(1 - \frac{D_V - \min (D_V)}{\max (D_V) - \min (D_V)}\right) - 1\|_1$$  \quad (3)

The above formulation affords two benefits, (1) since we use similarity scores to obtain $z'_V$, without involving ground truth, it enables $L_{\text{inter}}$ to guide the network to focus on a better video region candidate based on what the model has learned over training in case the ground truth for a certain sample $K$ is noisy. (2) the Gumbel_Softmax function keeps the weighted proposal differentiable to allow influencing the features from the entire untrimmed video.

#### Inter-Video-Text Alignment Loss

While $L_{\text{inter}}$ enforces a better alignment of the video with the query, it may still cause the model to not focus on the most representative region of the video. For this, we need a second alignment loss to increase the margin between the video regions that align better with the query and those that align poorly. Inspired by weakly-supervised grounding methods [53], we define Intra-Video-Text Alignment Loss $L_{\text{intra}}$ as,
\[ \mathcal{L}_{\text{intra}} = - \sum_{(z_V^f, z_Q)} \log \left( \frac{\exp(z_Q \cdot z_V^f / \tau)}{\sum_{z} \exp(z_Q \cdot z / \tau)} \right) \]

where \( z_V^f \) and \( z_Q \) are video features obtained by averaging \( P \) and \( N \) are positive and negative proposal sets containing the top-\( pr \) proposals with the highest and the lowest video-text similarity scores respectively (Fig 2e).

We optimize ProTéGé using a combination of the above three losses defined as \( \mathcal{L}_{\text{VTG}} = \mathcal{L}_{\text{loc}} + w_1 \cdot \mathcal{L}_{\text{inter}} + w_2 \cdot \mathcal{L}_{\text{intra}} \) where \( w_1 \) and \( w_2 \) are loss coefficients.

4. Experiments

Datasets. We evaluate ProTéGé on common VTG datasets, Charades-STA [8] and ActivityNet-Captions [16], using their standard splits. The former is built on the Charades dataset containing 9,848 videos of daily indoor activity scenarios with 12,408 and 3,720 video-sentence pairs in the training and test set, respectively. The latter is built on ActivityNet v1.3, which contains 20k YouTube videos, with 37,417, 17,505, and 17,031 video-sentence pairs in training and test set, respectively. The latter is built on the Charades-STA and ActivityNet-Captions respectively. From Table 1a, we can observe that compared to Swin-B and Swin-T video backbones which are pretrained on trimmed videos, ProTéGé, which is pretrained on untrimmed videos, can achieve significantly high improvement of 7.15%/5.6% and 4.20%/4.71% respectively on R@0.5/R@0.7 metrics on Charades-STA. We also outperform similarly on ActivityNet-Captions as shown in Table 1b. Moreover, we compare ProTéGé using Swin-T backbone with existing methods doing pretraining on trimmed video-text pairs with backbones of comparable size (Row 1-5,10 in Table 1a and Row 1-5,9 in Table 1b). ProTéGé significantly outperforms all baselines by at least 6.29%/5.13% and 1.02%/1.75% on R@0.5/R@0.7 metrics on Charades-STA and ActivityNet-Captions respectively. Further, ProTéGé using Swin-B outperforms LocVTP of comparable size by 9.66%/4.08% on R@0.5/R@0.7 metrics. We cannot compare ProTéGé with LocVTP on ActivityNet-Caption as LocVTP seems to have val.1 split in the training set which makes test videos part of training and artificially elevates the scores.

These results validate the effectiveness of ProTéGé and the significance of pretraining on untrimmed videos via VTG in achieving higher performance on downstream VTG in a supervised setting. Moreover, we achieve a higher performance of 4.17%/2.00% using Swin-B than Swin-T which also highlights the capability of ProTéGé to scale with more data and network parameters.

4.2. Weakly-supervised Video Temporal Grounding

We assess the effectiveness of the pretrained features from ProTéGé on weakly-supervised VTG task. This task assumes that while the query is available during training, we do not have access to its location in the video in the form of a start-end timestamp. We choose CPL on Charades-STA and CNM on ActivityNet-Captions which achieve state-of-the-art (SoTA) performance on the respective datasets. Following CPL and CNM, we report Top-1 Recall at tIoU threshold 0.3, 0.5, 0.7 for Charades-STA and 0.1, 0.3, 0.5 for ActivityNet-Captions as well as mIoU metric.

Table 1c tabulates the results for Charades-STA. We report results of CPL on Swin-T and Swin-B for a fair comparison. From the table, we can observe that using Swin-T and Swin-B with CPL, our method pretrained on untrimmed videos outperforms corresponding video backbones pretrained on trimmed videos by 1.58% and 0.98% respectively on mIoU metric. ProTéGé exceeds all other baselines by at least 3.17% and 4.82% using Swin-T and Swin-B respectively, again validating that it can achieve even higher performance with larger-scale backbones. Table 1c shows the comparison on ActivityNet-Captions. Comparing ProTéGé...
Table 1. Fully-supervised Video Temporal Grounding using 2D-TAN on (a) Charades-STA and (b) ActivityNet-Captions. Weakly-supervised Video Temporal Grounding on (c) Charades-STA and (e) ActivityNet-Captions. (d) Zero-shot Video Temporal Grounding on Charades-STA and Activity-Caption. Using pretrained features from ProTéGé shows significant improvements over features from all existing trimmed pretrained backbones over all benchmarks. *reported in paper, †obtained via official model weights

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>R@0.3</th>
<th>R@0.5</th>
<th>R@0.7</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Charades-STA</td>
<td>PSVL [33]</td>
<td>46.47</td>
<td>31.29</td>
<td>14.17</td>
<td>31.24</td>
</tr>
<tr>
<td></td>
<td>PSVL [33]</td>
<td>46.63</td>
<td>30.84</td>
<td>13.57</td>
<td>30.09</td>
</tr>
<tr>
<td></td>
<td>PSVL + Ours</td>
<td>46.79</td>
<td>31.84</td>
<td>17.51</td>
<td>31.25</td>
</tr>
<tr>
<td></td>
<td>PSVL + [CLIP] [33]</td>
<td>44.74</td>
<td>30.08</td>
<td>14.74</td>
<td>29.62</td>
</tr>
<tr>
<td>(b) ActivityNet-Captions</td>
<td>PSVL [33]</td>
<td>43.03</td>
<td>25.17</td>
<td>10.98</td>
<td>30.78</td>
</tr>
<tr>
<td></td>
<td>PSVL + Ours</td>
<td>45.02</td>
<td>27.85</td>
<td>14.89</td>
<td>33.04</td>
</tr>
</tbody>
</table>

4.3. Zero-shot Video Temporal Grounding

To further demonstrate ProTéGé’s ability to be effective on VTG tasks with varying degrees of supervision, we show the performance of using pretrained features from ProTéGé on zero-shot video temporal grounding. This downstream task assumes that during training, neither the text query nor the corresponding start-end timestamp is available. We demonstrate this task using PSVL on Charades-STA and ActivityNet-Captions and report results using Top-1 Recall at tIoU thresholds of 0.3, 0.5, and 0.7.

Table 1d shows the results on Charades-STA and ActivityNet-Captions respectively. For a fair comparison, we compare with results from PSVL obtained via the officially provided model weights but we also provide the results reported by PSVL in the paper for reference. We use Swin-T for ProTéGé’s experiments as it is a comparable backbone. We can observe that ProTéGé significantly outperforms PSVL with a mIoU improvement of 1.14% on Charades-STA and 2.26% on ActivityNet-Captions. Improvements from ProTéGé on this benchmark further establish the usefulness of performing untrimmed pretraining via VTG to benefit all types of downstream VTG tasks.

4.4. Ablation Study

To evaluate the contribution of each novel component of ProTéGé, we conduct an ablation study. We experiment on Charades-STA using Swin-B backbone on fully-supervised VTG and report results in Table 2a. We first observe that removing the localization loss, \( \mathcal{L}_{loc} \), from our pretraining objective leads to a drop in performance by 3.86%/3.78% on the R@0.5/R@0.7 metric. This shows that even though the video-text correlations are noisy, it is still important to leverage the ground truth \( SE \) via a combination of \( \mathcal{L}_{CE} \) and \( \mathcal{L}_{dist} \) to perform effective pretraining. Next, we observe that if we instead remove the video-text alignment loss, \( \mathcal{L}_{inter} \) and \( \mathcal{L}_{intra} \), from the pretraining objective, it reduces the performance by 1.63%/0.79%. This helps to validate that without the alignment losses and fully relying on the localization loss is not optimal as it makes the module more vulnerable to the noisy video-text correlations due to increased dependence on the ground truth \( SE \).

We also explore the contribution of our novel grounding module, VT-SGM, towards better pretraining. We con-
duct an experiment where we replace our 2D proposal grid with a regression layer that directly predicts the start and end timestamp for grounding. We find this setting to perform significantly worse with a drop of 4.9%/5.16%. This proves that given the imperfect video-text correlations in the pretraining dataset, direct regression is prone to noise and our VT-SGM allows for a softer localization to learn better grounding-oriented feature representations. We further validate the importance of pretraining on untrimmed videos where we directly perform video-text alignment using our model backbone. This performs 5.91%/4.91% worse than ProTéGé, highlighting the merit of pretraining on untrimmed videos.

Table 3. Analysis of different loss functions showing that each loss contributes significantly towards the optimal performance.

<table>
<thead>
<tr>
<th>Method</th>
<th>R@0.5</th>
<th>R@0.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>53.26</td>
<td>30.38</td>
</tr>
<tr>
<td>Localization loss</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- w/o $L_{CE}$</td>
<td>51.18</td>
<td>27.33</td>
</tr>
<tr>
<td>- w/o $L_{dist}$</td>
<td>51.86</td>
<td>28.35</td>
</tr>
<tr>
<td>- w/o $L_{CE} + L_{dist}$</td>
<td>49.40</td>
<td>26.60</td>
</tr>
<tr>
<td>Alignment loss</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- w/o $L_{intra}$</td>
<td>52.45</td>
<td>29.22</td>
</tr>
<tr>
<td>- w/o $L_{inter}$</td>
<td>52.39</td>
<td>28.88</td>
</tr>
<tr>
<td>- w/o $L_{intra} + L_{inter}$</td>
<td>51.63</td>
<td>29.59</td>
</tr>
</tbody>
</table>

4.5. Discussion

We assess the different components of ProTéGé in detail to understand their influence on pretraining and downstream performance. We experiment on Charades-STA using the Swin-B backbone on fully-supervised VTG.

Pretraining Objective. Table 3 reports the performance for different combinations of losses in our pretraining objective. For localization loss without cross-entropy, $L_{CE}$, the performance drops by 2.08%/3.05% while removing the distance loss, $L_{dist}$, leads to a reduction of 1.4%/2.03%. This highlights that both $L_{CE}$ and $L_{dist}$ contribute significantly in the pretraining objective. Either soft-localization via $L_{dist}$ or hard-localization via $L_{CE}$ alone causes the model to under-utilize the available ground truth $SE$ information in the synthesized VTG pretraining dataset. We similarly find removing the inter-alignment loss, $L_{inter}$ to reduce performance by 0.87%/1.5% and removing intra-alignment loss, $L_{intra}$ to reduce performance by 0.81%/1.16%. This validates that both video-text alignment losses contribute to the best performance of ProTéGé. Both losses leverage representation learning to make the model robust to noisy ground truth; $L_{inter}$ leverages it via alignment and $L_{intra}$ uses it to make features within a video more discriminative.

Top-pr proposals in $L_{intra}$. Table 2b further shows the results on selecting a different number of proposals as part of

4.5. Discussion

We assess the different components of ProTéGé in detail to understand their influence on pretraining and downstream performance. We experiment on Charades-STA using the Swin-B backbone on fully-supervised VTG.

Pretraining Objective. Table 3 reports the performance for different combinations of losses in our pretraining objective. For localization loss without cross-entropy, $L_{CE}$, the performance drops by 2.08%/3.05% while removing the distance loss, $L_{dist}$, leads to a reduction of 1.4%/2.03%. This highlights that both $L_{CE}$ and $L_{dist}$ contribute significantly in the pretraining objective. Either soft-localization via $L_{dist}$ or hard-localization via $L_{CE}$ alone causes the model to under-utilize the available ground truth $SE$ information in the synthesized VTG pretraining dataset. We similarly find removing the inter-alignment loss, $L_{inter}$ to reduce performance by 0.87%/1.5% and removing intra-alignment loss, $L_{intra}$ to reduce performance by 0.81%/1.16%. This validates that both video-text alignment losses contribute to the best performance of ProTéGé. Both losses leverage representation learning to make the model robust to noisy ground truth; $L_{inter}$ leverages it via alignment and $L_{intra}$ uses it to make features within a video more discriminative.

Top-pr proposals in $L_{intra}$. Table 2b further shows the results on selecting a different number of proposals as part of

4.5. Discussion

We assess the different components of ProTéGé in detail to understand their influence on pretraining and downstream performance. We experiment on Charades-STA using the Swin-B backbone on fully-supervised VTG.

Pretraining Objective. Table 3 reports the performance for different combinations of losses in our pretraining objective. For localization loss without cross-entropy, $L_{CE}$, the performance drops by 2.08%/3.05% while removing the distance loss, $L_{dist}$, leads to a reduction of 1.4%/2.03%. This highlights that both $L_{CE}$ and $L_{dist}$ contribute significantly in the pretraining objective. Either soft-localization via $L_{dist}$ or hard-localization via $L_{CE}$ alone causes the model to under-utilize the available ground truth $SE$ information in the synthesized VTG pretraining dataset. We similarly find removing the inter-alignment loss, $L_{inter}$ to reduce performance by 0.87%/1.5% and removing intra-alignment loss, $L_{intra}$ to reduce performance by 0.81%/1.16%. This validates that both video-text alignment losses contribute to the best performance of ProTéGé. Both losses leverage representation learning to make the model robust to noisy ground truth; $L_{inter}$ leverages it via alignment and $L_{intra}$ uses it to make features within a video more discriminative.

Top-pr proposals in $L_{intra}$. Table 2b further shows the results on selecting a different number of proposals as part of

5. Conclusion

We present ProTéGé as the first method to bridge the gap between pretraining and downstream VTG by pretraining on untrimmed videos via VTG. To do so, ProTéGé first synthesizes a VTG pretraining dataset from large-scale video dataset HowTo100M with noisy video-text pairs using aggregated subtitles and then performs pretraining via a novel Video-Text Similarity-based Grounding Module (VT-SGM) and pretraining objective comprising a localization loss and pretraining objective.

Acknowledgements. At MSU, Lan Wang is supported in part by award 60NANB18D210 by NIST, U.S. Dept. of Commerce.
References


[27] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou.


[51] Songyang Zhang, Houwen Peng, Jialong Fu, and Liebo Luo. Learning 2d temporal adjacent networks for moment local-


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 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>
<thead>
<tr>
<th>Max duration of $Q$ (s)</th>
<th>R@0.5</th>
<th>R@0.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>52.59</td>
<td>27.48</td>
</tr>
<tr>
<td>20</td>
<td>50.19</td>
<td>26.83</td>
</tr>
<tr>
<td>50</td>
<td>53.26</td>
<td>30.38</td>
</tr>
<tr>
<td>100</td>
<td>48.27</td>
<td>27.73</td>
</tr>
</tbody>
</table>

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

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>
<thead>
<tr>
<th>Max video clips $M$</th>
<th>R@0.5</th>
<th>R@0.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>46.19</td>
<td>25.14</td>
</tr>
<tr>
<td>30</td>
<td>47.98</td>
<td>25.59</td>
</tr>
<tr>
<td>60</td>
<td>53.26</td>
<td>30.38</td>
</tr>
<tr>
<td>120</td>
<td>53.24</td>
<td>30.60</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Batch size $B$</th>
<th>R@0.5</th>
<th>R@0.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>512</td>
<td>50.67</td>
<td>29.11</td>
</tr>
<tr>
<td>1024</td>
<td>52.51</td>
<td>30.86</td>
</tr>
<tr>
<td>2048</td>
<td>53.26</td>
<td>30.38</td>
</tr>
<tr>
<td>4096</td>
<td>51.74</td>
<td>28.69</td>
</tr>
</tbody>
</table>

Table 3. Effect of batch size $B$ during pretraining.

* 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 \( \mathcal{L}_{\text{inter}} \) 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 [5].

**Evaluation on VT-SGM.** The proposed grounding module (VT-SGM) is directly incorporated into pretraining to leverage untrimmed videos for VTG. While downstream VTG methods like 2D-TAN 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 original 2D-TAN and finding that it leads to 2.7%/4.0% lower R@0.5/R@0.7 on Charades and 1.8%/1.8% lower R@0.3/R@0.5 on TACoS, as shown in Table 4, further showing the benefit of the novel design.

<table>
<thead>
<tr>
<th>Table 4. Evaluation on VT-SGM.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Charades-STA</strong></td>
</tr>
<tr>
<td><strong>Grounding Module</strong></td>
</tr>
<tr>
<td><strong>VT-SGM</strong></td>
</tr>
<tr>
<td><strong>2D-TAN</strong></td>
</tr>
</tbody>
</table>

**Evaluation on more datasets.** We show video temporal grounding evaluation on more datasets. Table 6 and 5 shows results on TACoS and QVHighlights using ProTéGé with Moment-DETR and 2D-TAN as downstream methods. It can be observed 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/R@0.5 on TACoS vs. trimmed pretraining (Row 2) using the same extra data, which empirically validates the our untrimmed pretraining model. Moreover, comparing with the baselines, ProTéGé also shows superior performance.

<table>
<thead>
<tr>
<th>Table 5. Evaluation on QVHighlights.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Method</strong></td>
</tr>
<tr>
<td>Moment-DETR [1]</td>
</tr>
<tr>
<td>ProTéGé w/o Untrimmed</td>
</tr>
<tr>
<td>ProTéGé</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 6. Evaluation on TACoS.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Method</strong></td>
</tr>
<tr>
<td>LocVTP[5]</td>
</tr>
<tr>
<td>ProTéGé w/o Untrimmed</td>
</tr>
<tr>
<td>ProTéGé</td>
</tr>
</tbody>
</table>

**Evaluation on VidSitu.** To demonstrate the effectiveness of the proposed method on understanding movie data, we further evaluate ProTéGé on VidSitu, 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 as the downstream method. Table 7 shows MacroAveraged Accuracy on validation set. ProTéGé exceeds Vid TxEnc and trimmed pretraining, furthur validating its generalization ability on movie data and complex situation understanding task.

<table>
<thead>
<tr>
<th>Table 7. Evaluation on VidSitu.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Method</strong></td>
</tr>
<tr>
<td>Vid TxEnc</td>
</tr>
<tr>
<td>ProTéGé w/o Untrimmed</td>
</tr>
<tr>
<td>ProTéGé</td>
</tr>
</tbody>
</table>

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 [6] 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 [3] pretrained on Kinetics-400 and Swin-B [3] pretrained on Kinetics-600 as the frozen trimmed video encoders, $f^v$. 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 [4] implementation of RoBERTa-base [2] for the frozen text encoder $f^q$. 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 $224 \times 224$ 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


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)

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

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

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

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.