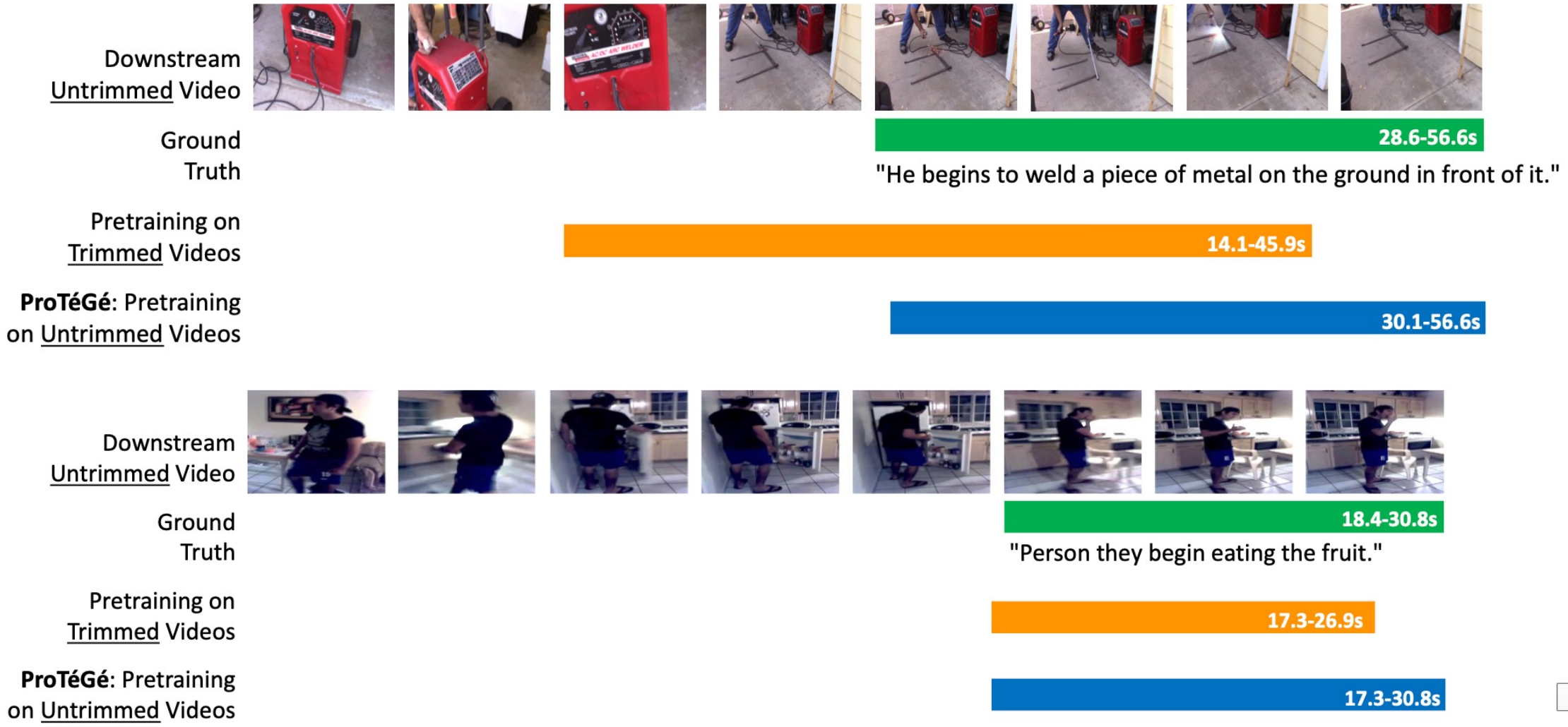
ProTéGé: Untrimmed Pretraining for Video Temporal Grounding by Video Temporal Grounding

Lan Wang[*2]**, Gaurav Mittal**[*1]**,** Sandra Sajeev[1], Ye Yu[1], Matthew Hall[1], Vishnu Naresh Boddeti[2], Mei Chen[1]

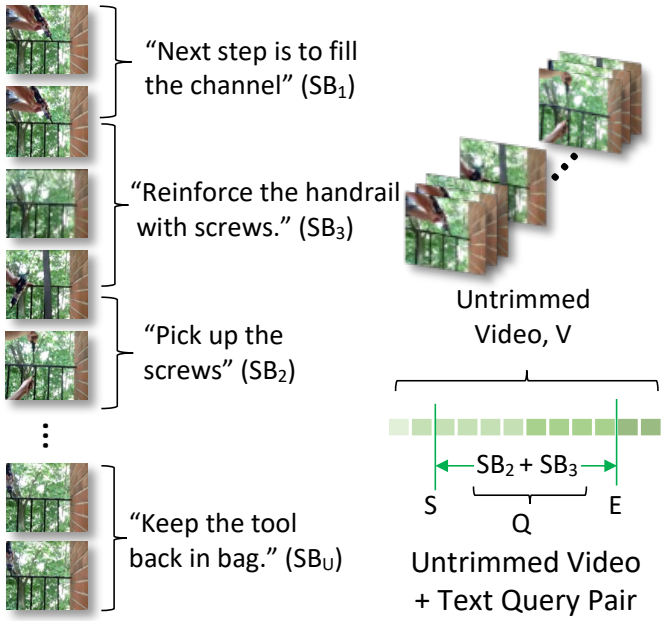
[1]Microsoft [2]Michigan State University

TUE-PM-234

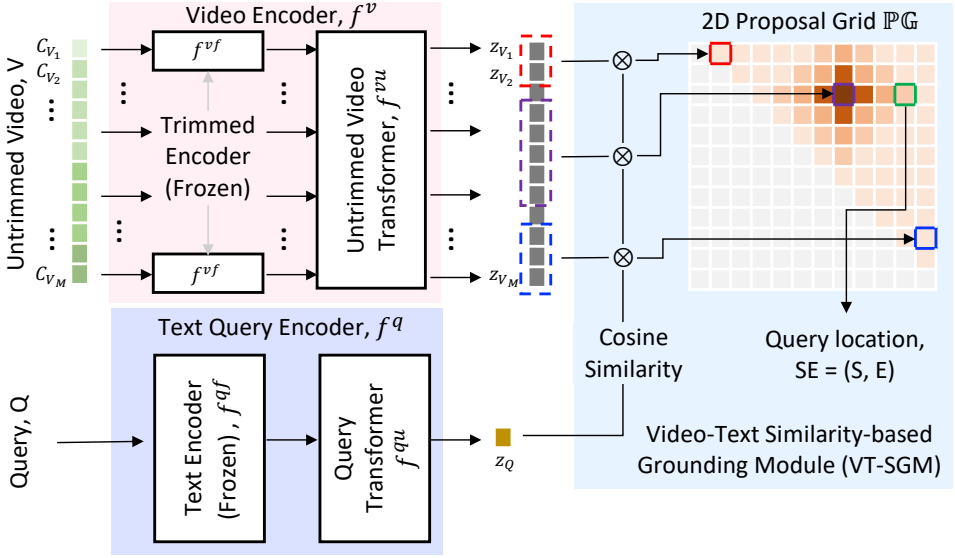
ProTéGé bridges the gap between video-text pretraining and downstream video temporal grounding (VTG) tasks by formulating pretraining as a VTG task over untrimmed videos



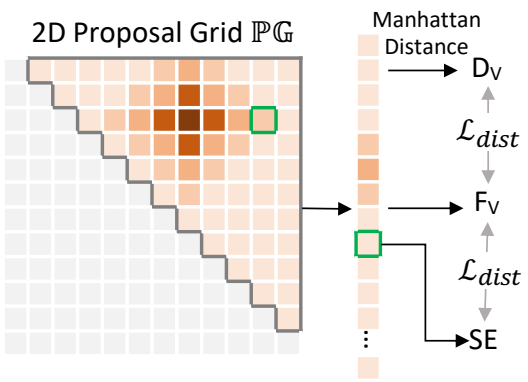
Create VTG Pretraining Dataset from HowTo100M



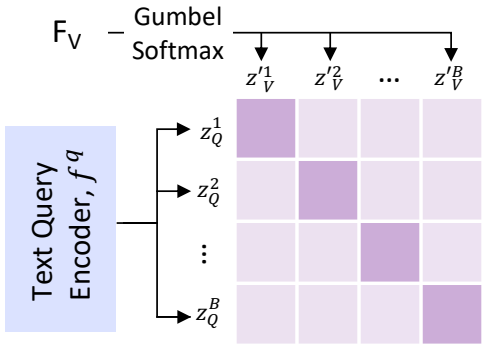
Encoding untrimmed video-text pairs into a 2D Proposal Grid of cosine similarity scores



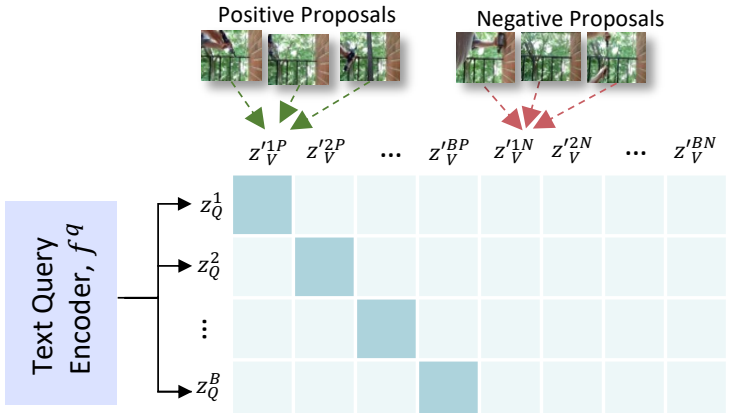
Pretraining Objective (Localization + Video-Text Alignment Losses)



Localization Loss, $\mathcal{L}_{loc} = \mathcal{L}_{CE} + \mathcal{L}_{dist}$

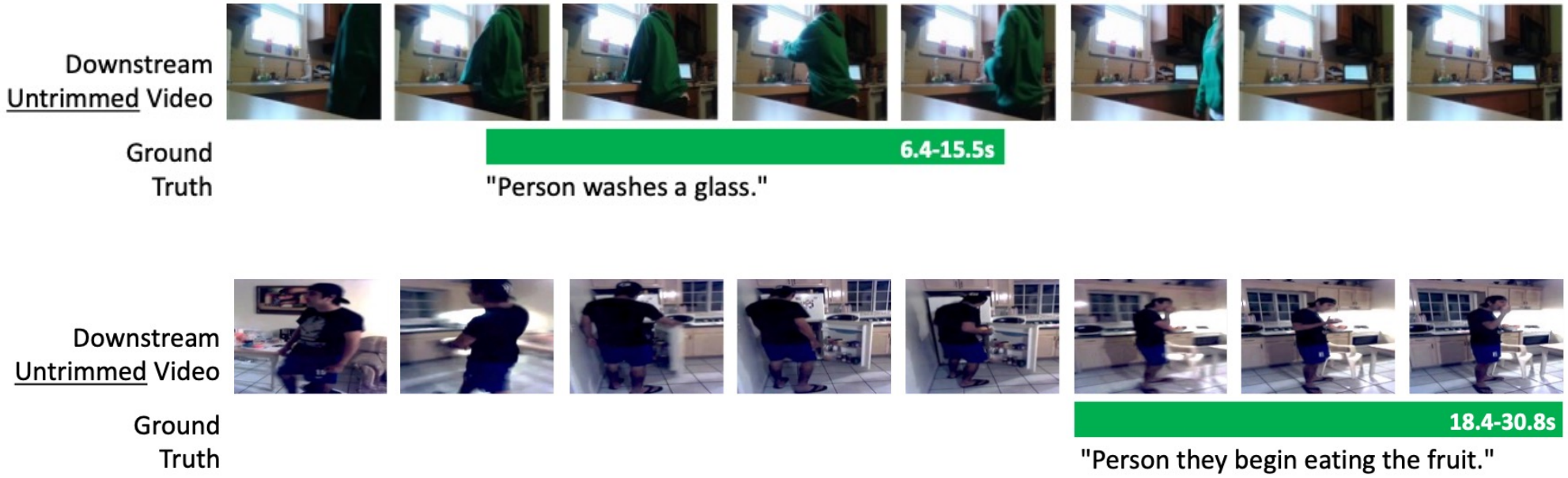


Inter Video-Text Alignment Loss



Intra Video-Text Alignment Loss

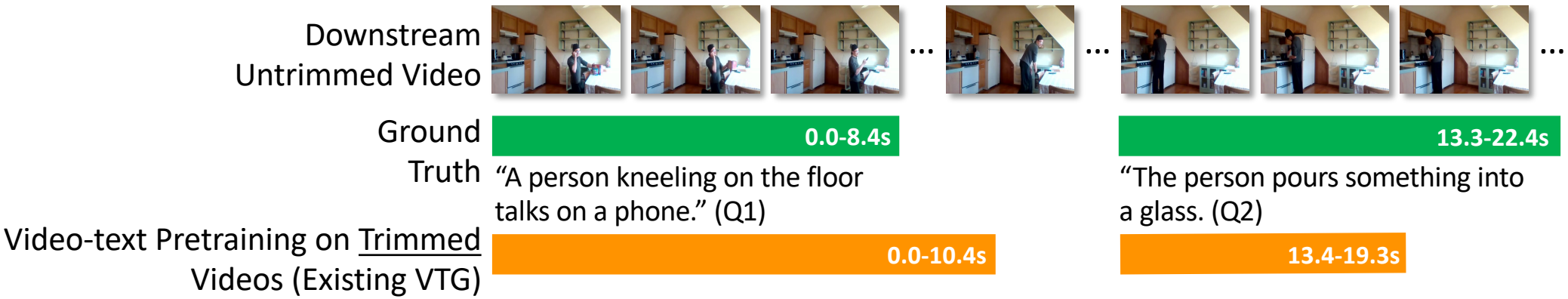
Video Temporal Grounding (VTG)



Given an untrimmed video and text prompt, localize which part of the video is best associated with the text prompt.

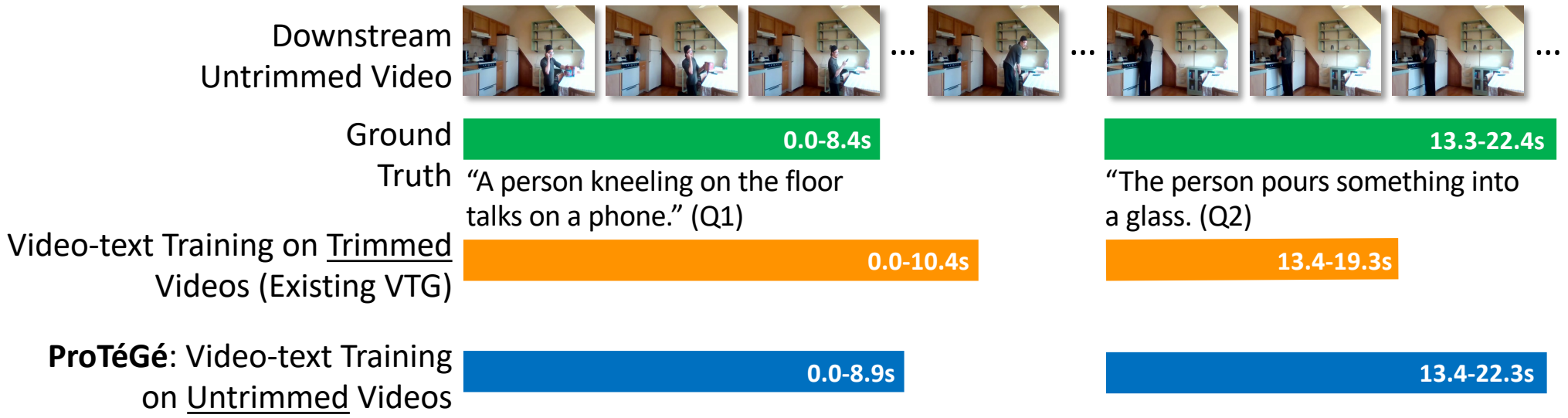
Useful for video question answering and searching events in videos.

Limitations of Existing VTG Methods



- ⚠ Rely on video backbones pretrained on trimmed videos
- ⚠ Insensitive to temporal boundaries
- ⚠ No explicit ability to localize

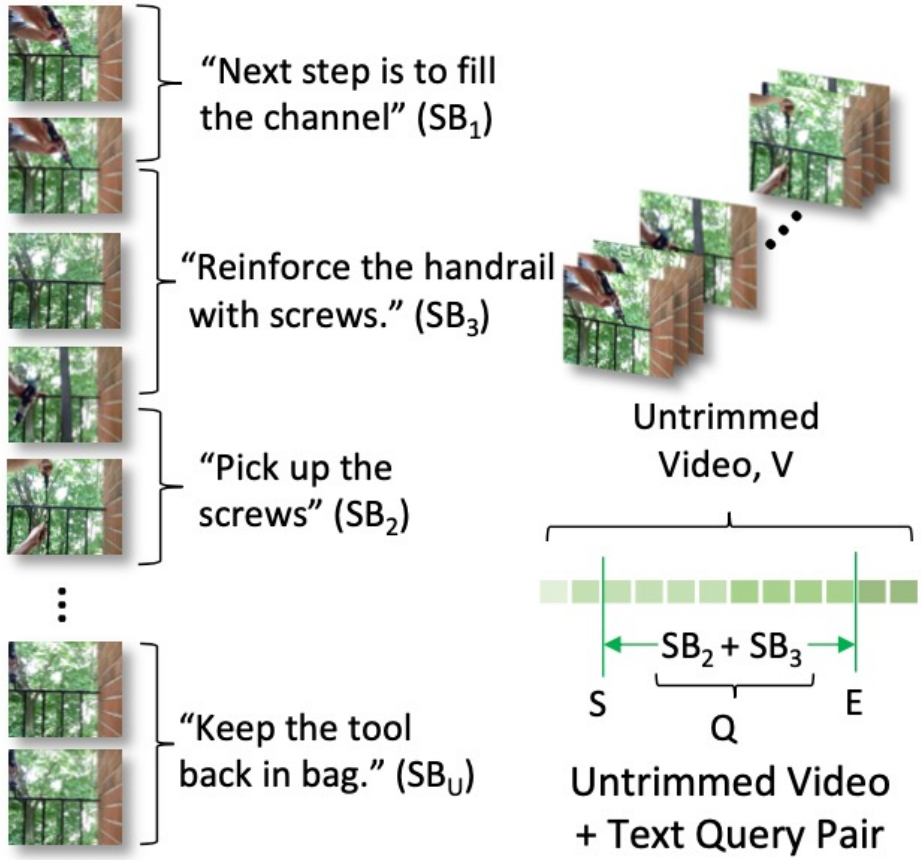
ProTéGé: Untrimmed VTG Pretraining for VTG



👍 ProTéGé performs untrimmed video-text pretraining as a VTG task to enable fine-grained understanding of temporal boundaries

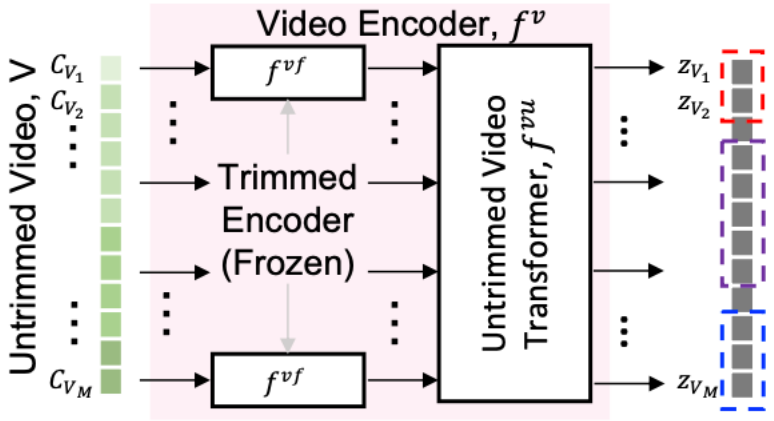
Synthesizing VTG Dataset for Untrimmed Pretraining

- HowTo100M → VTG-based untrimmed pretraining in ProTéGé
- HowTo100M: 1.22M untrimmed videos with autogenerated speech-to-text subtitles (SB).
- **Aggregated Subtitles:** Concatenate one or more captions together to form text prompt for VTG
- **Untrimmed Video-Text Query pair:** From a randomly selected video, create an aggregated subtitle and randomly subsample video segment around it



Model Overview

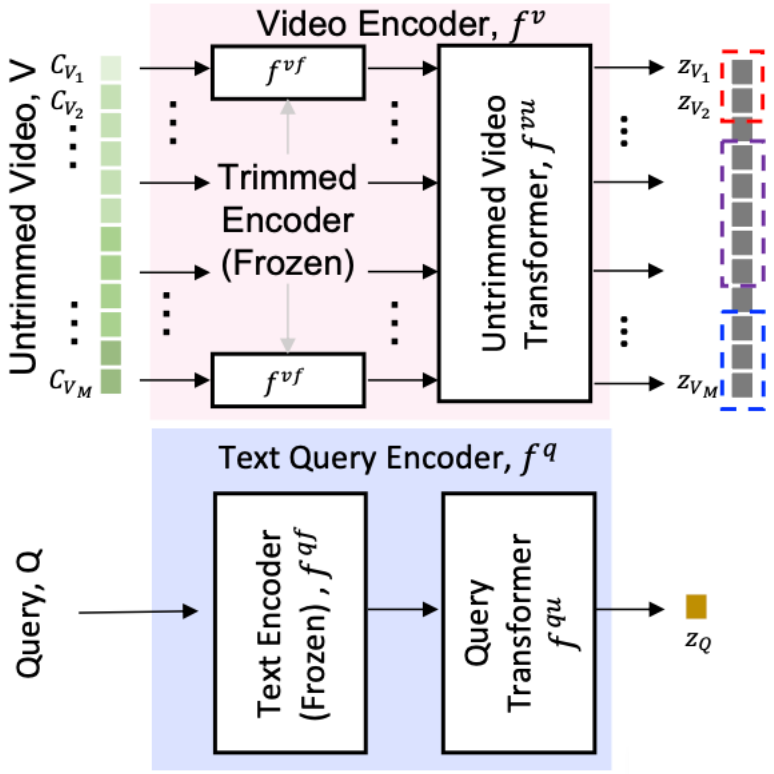
- Untrimmed video split into segments, encoded by frozen trimmed encoder, then by learnable untrimmed video transformer



Model Overview

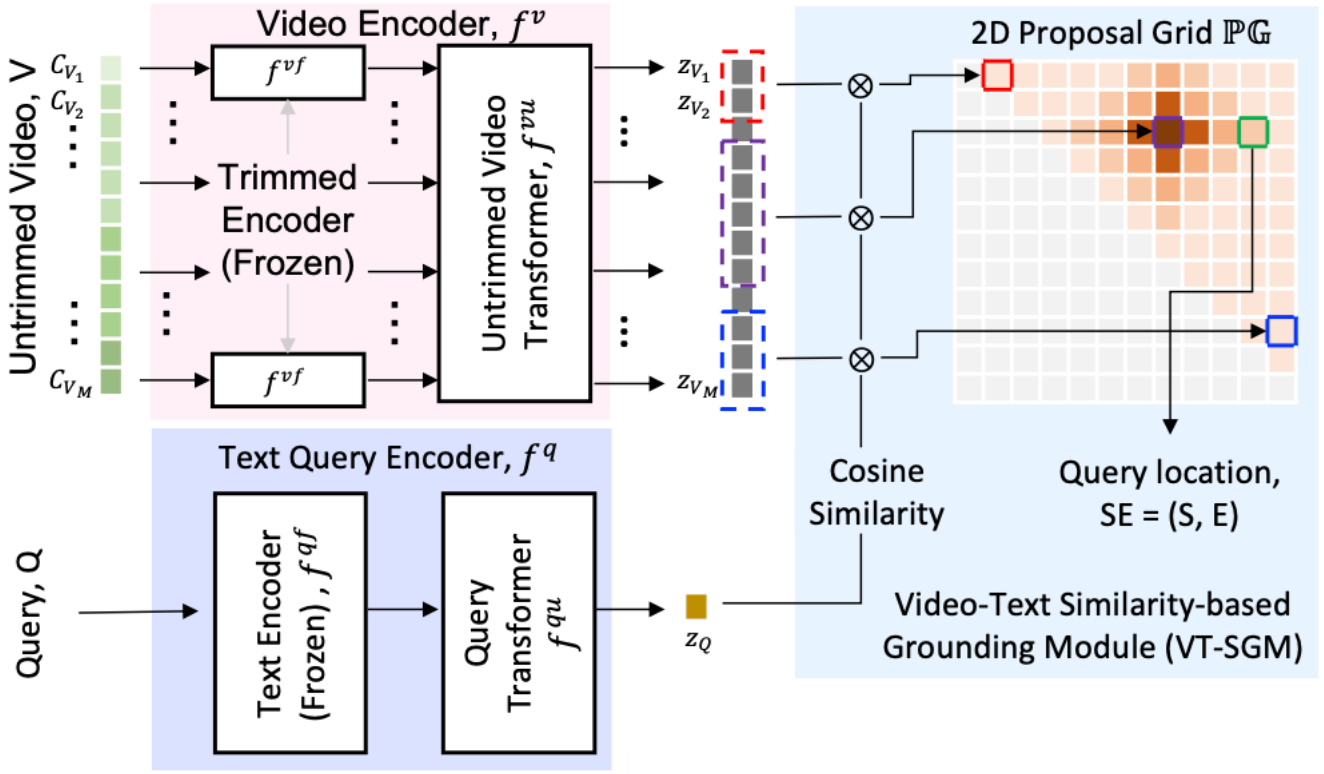
- Untrimmed video split into segments, encoded by frozen trimmed encoder, then by learnable untrimmed video transformer

- Text query first encoded by frozen text encoder, then by learnable query transformer



Model Overview

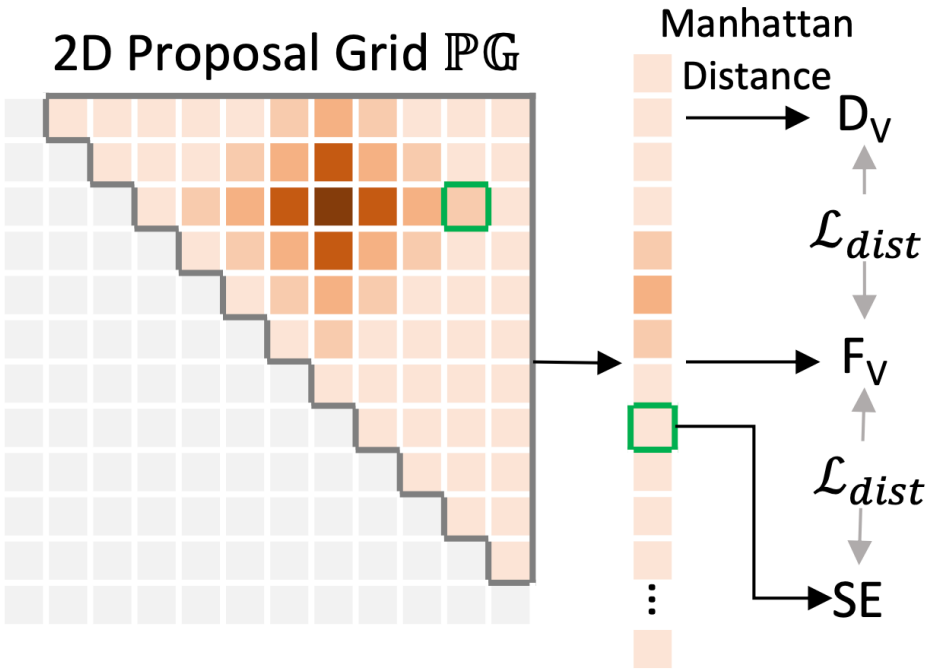
- Untrimmed video split into segments, encoded by frozen trimmed encoder, then by learnable untrimmed video transformer
- Text query first encoded by frozen text encoder, then by learnable query transformer
- **Video-Text Similarity-based Grounding Module (VT-SGM):** Outputs grid of cosine similarity between text features and all possible sub-sequences of video features



Pretraining Objective: Localization Loss

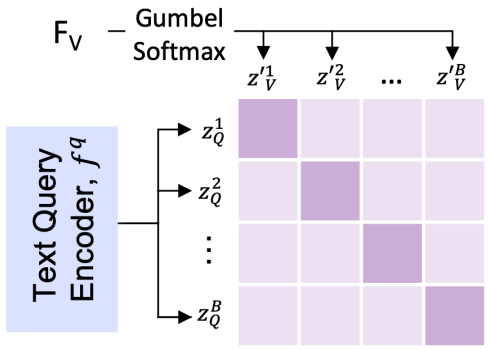
- **Distance-based soft-localization loss, $\mathcal{L}_{dist}$:** Manhattan distance between cosine similarity scores and ground truth SE (where text localizes in video)

- **Cross-Entropy Loss , $\mathcal{L}_{ce}$:** Between cosine similarity scores and ground truth SE

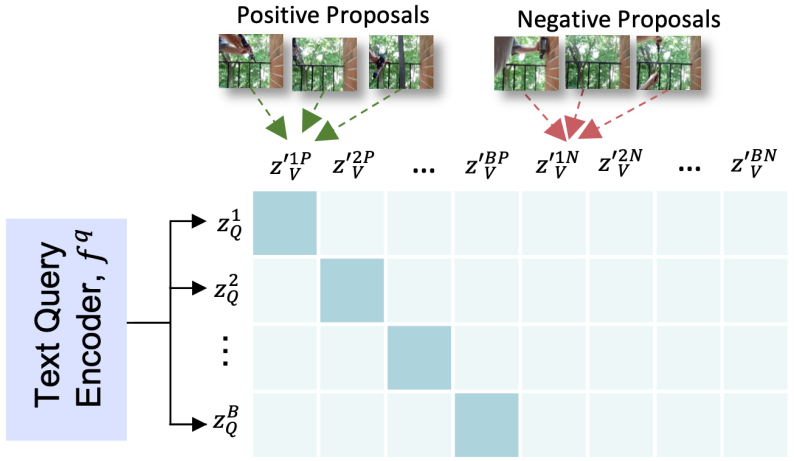


Localization Loss, $\mathcal{L}_{loc} = \mathcal{L}_{CE} + \mathcal{L}_{dist}$

Pretraining Objective: Video-Text Alignment Loss



Inter Video-Text Alignment Loss



Intra Video-Text Alignment Loss

Inter-Video-Text Alignment Loss, $\mathcal{L}_{inter}$:
Maximizes alignment of text query Q with video V among alignments of Q with all videos in training batch.

Intra-Video-Text Alignment Loss, $\mathcal{L}_{intra}$:
Increases the margin between video regions aligning better with text query Q (+ve proposals) and those that align poorly (-ve proposals).

Experiment Setup

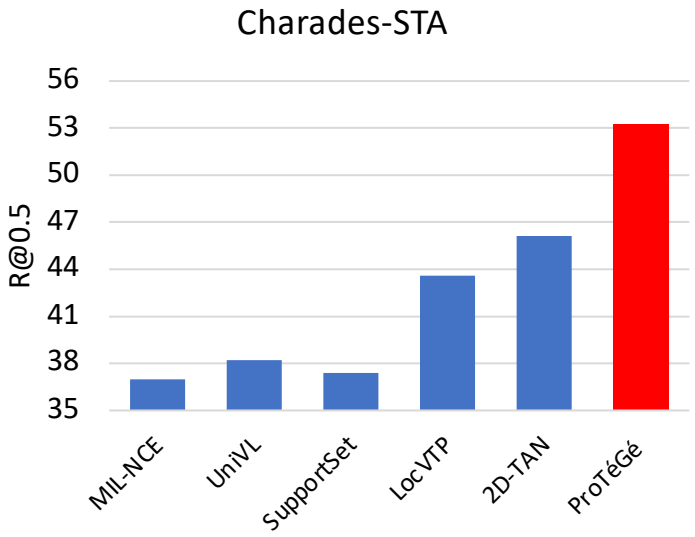
Downstream Video-Temporal Grounding Tasks

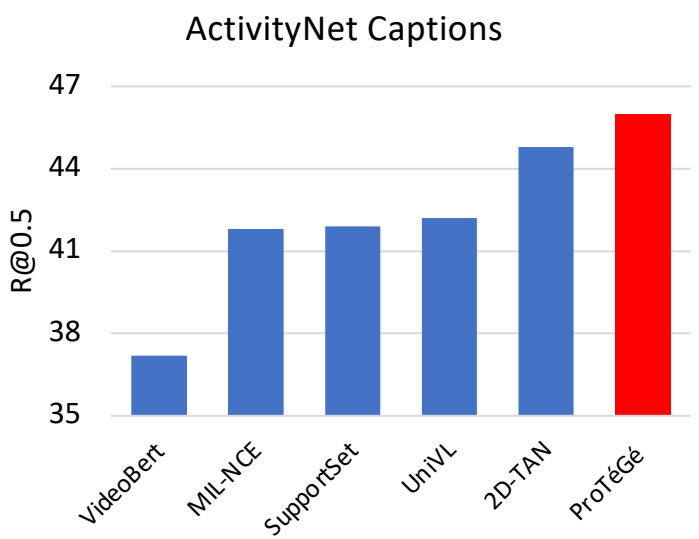
- Fully-supervised
- Weakly-supervised
- Zero-shot

Datasets

- Charades-STA
- ActivityNet-Captions

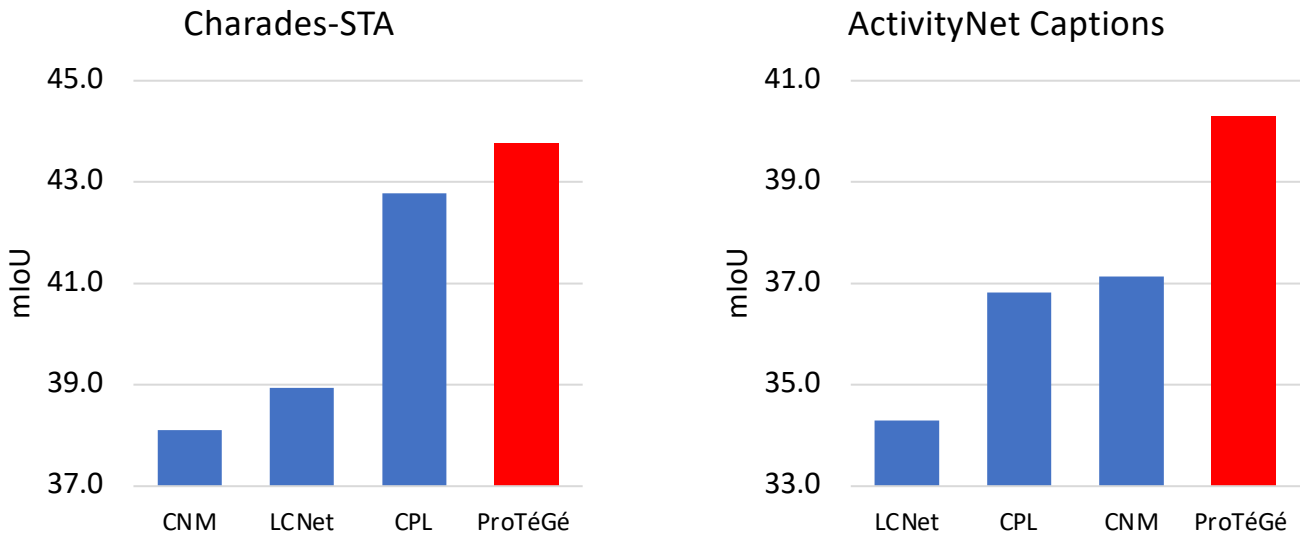
ProTéGé: Fully-Supervised Video Temporal Grounding





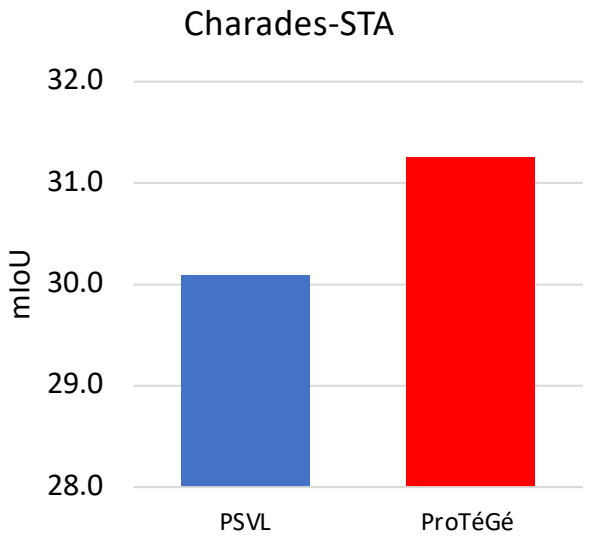
ProTéGé outperforms existing methods when finetuned on downstream fully-supervised VTG task

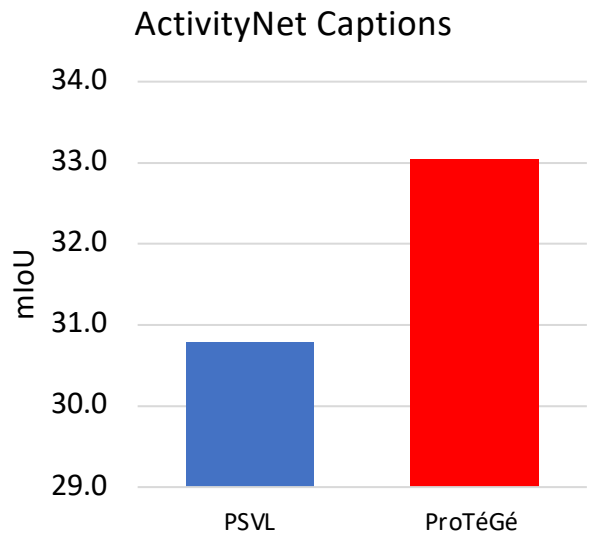
ProTéGé: Weakly-Supervised Video Temporal Grounding



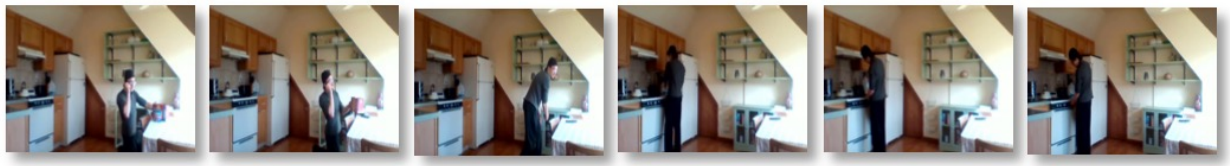
ProTéGé outperforms existing methods when finetuned on downstream weakly-supervised VTG task

ProTéGé: Zero-shot Video Temporal Grounding

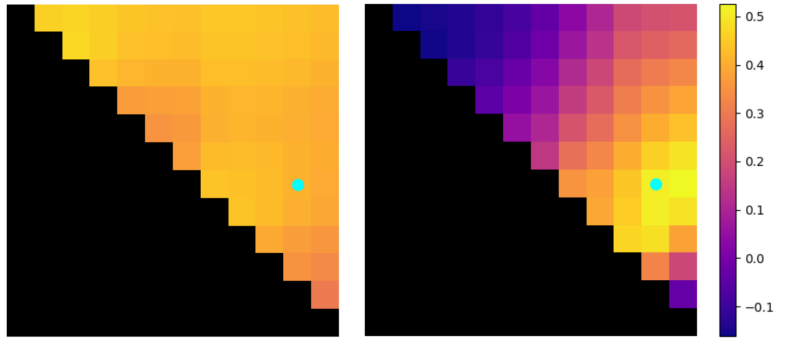




Visualization on unseen videos



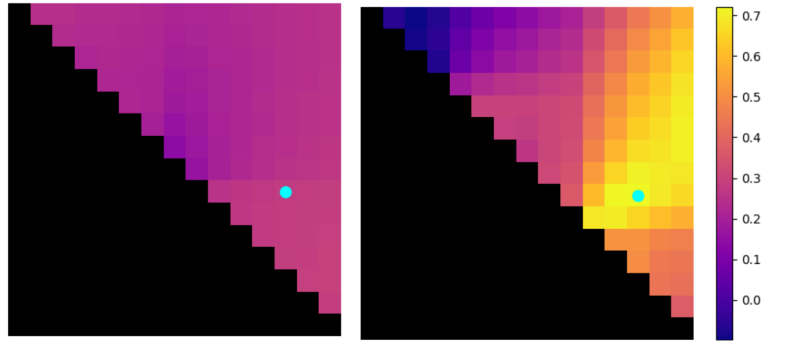
Video 1 : “The person pours something into a glass.” (Start - 13.3s, End - 22.4s, Duration - 23.83s)



Trimmed pretraining Protégé

Video 1

Video 2 : “One person uses a camera to take a picture.” (Start - 17.5s, End - 25.8s, Duration - 32.71s)



Trimmed pretraining Protégé

Video 2

Conclusion

- **ProTéGé:** First pretraining method formulated as a VTG task to bridge the gap between pretraining and downstream VTG tasks in untrimmed videos.

- Significant performance improvement across all downstream VTG benchmarks across different levels of supervision.

- **Hope to see you at our poster TUE-PM-234!** 😊