Zero-shot Natural Language Video Localization

Jinwoo Nam[1,*] Daechul Ahn[1,*] Dongyeop Kang[2,§] Seong Jong Ha[3] Jonghyun Choi[1,†]
[1]GIST, South Korea [2]UC Berkeley [3]Vision AI Lab, AI Center, NCSOFT
{skaws2003, daechulahn}@gm.gist.ac.kr, dongyeopk@berkeley.edu, seongjongha@ncsoft.com, jhc@gist.ac.kr

Abstract

Understanding videos to localize moments with natural language often requires large expensive annotated video regions paired with language queries. To eliminate the annotation costs, we make a first attempt to train a natural language video localization model in zero-shot manner. Inspired by unsupervised image captioning setup, we merely require random text corpora, unlabeled video collections, and an off-the-shelf object detector to train a model. With the unpaired *data, we propose to generate pseudo-supervision of candidate temporal regions and corresponding query sentences, and develop a simple NLVL model to train with the pseudo-supervision. Our empirical validations show that the proposed pseudo-supervised method outperforms several baseline approaches and a number of methods using stronger supervision on Charades-STA and ActivityNet-Captions.*

1. Introduction

On increasing demands of understanding videos to search with natural language queries, natural language video localization (NLVL) has been actively investigated in recent literature [19,35,39,43,44,57]. The task targets to localize a temporal moment in a video by a natural language query. In recent years, significant performance improvements on benchmark datasets has been made, facilitated by the advances on deep learning methods [19, 39, 43, 45] and massively annotated data [2, 19, 26, 37, 58].

As illustrated in Fig. 1-(a), the annotations consist of a temporal region in a video (start time, end time) and a corresponding query sentence. However, obtaining such paired annotation is laborious and expensive. To alleviate the annotation cost, a number of recent works addressed weakly-supervised setup of NLVL [12, 21, 33] which aims to localize a moment without the temporal alignment of given query sentence. Although it eliminates the annotation cost of specifying start and end points of the query sentence in video (illustrated in Fig. 1-(b)), the remaining cost of annotating natural language query is still considerable [15].

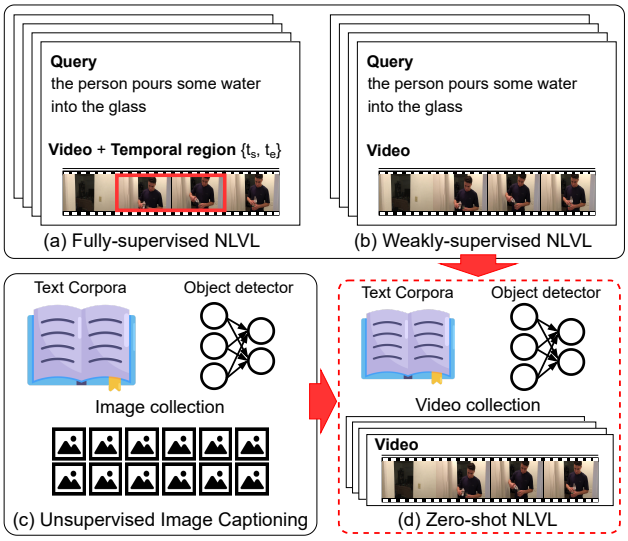

Figure 1: **Tasks with different levels of supervision.** (a) Supervised NLVL (queries and temporal regions on video) (b) Weakly-Supervised NLVL [21,33] (query on videos) (c) Unsupervised Image Captioning [15,29] (on images) (d) Proposed Zero-shot NLVL (on videos).

To avoid the costly annotations, we propose *zero-shot NLVL* (ZS-NLVL) task setup which aims to learn an NLVL model without any paired annotation, the first in the literature to our best knowledge. Inspired by [15, 29] addressing an image captioning task only with unpaired images, natural language corpora, and an object detector (Fig. 1-(c)), we propose to train an NLVL model by leveraging easily accessible unpaired data including videos, natural language corpora, and an *off-the-shelf* object detector, with no knowledge about video data to localize [3, 15]. We depict the given data for the zero-shot NLVL setup in Fig. 1-(d).

To address this task, we approach to generate *pseudo-supervision* of candidate temporal regions in video and corresponding sentences to train an NLVL model. The pseudo-supervision approach has several benefits as follows. First,

*: equal contribution. †: corresponding author. § now at U. of Minnesota, Twin Cities. **Code**: https://github.com/gistvision/PSVL

it provides interpretable resources (*i.e.*, generated regions and sentences) to train an NLVL model. Second, the pseudo-supervision can serve as initial annotation suggestions to human labelers to reduce the annotation cost or to accelerate the annotation process. Finally, the pseudo-supervision can be readily applicable to the existing 'fully supervised' NLVL models (Sec. 4.1.3).

Generating the pseudo supervision for NLVL involves two challenges: 1) finding meaningful temporal regions to be possibly queried and 2) obtaining corresponding query sentences for the temporal regions found. To find the possible temporal regions, we propose to cluster visual information (Sec. 3.1). Once we have the candidate (predicted) temporal regions, we obtain the corresponding (paired) query sentences. For that, we propose to find nouns visible in the frame by the off-the-shelf object detector and predict verbs that are likely appearing together with the detected objects by exploiting noun-verb statistical co-occurrence patterns from the language corpora (Sec. 3.2). We call the set of nouns and verbs as a *pseudo query*.

Since the pseudo query is structure-less unlike the natural language queries from the supervised data and not all the proposed event regions might be meaningful, we further propose a simple NLVL model which is better suited to such pseudo-supervision. We call this framework of training an NLVL model with the temporal region proposals and pseudo-query generation, as *Pseudo-Supervised Video Localization* or **PSVL** (Sec. 4.1).

Our empirical studies show that our PSVL exhibits competitive accuracy, sometimes outperforming the models with stronger supervision on widely used two benchmarks.

We summarize our contributions as follows:

- We propose the first zero-shot NLVL task.
- We propose an pseudo supervising framework (PSVL) to predict temporal event regions and corresponding query sentences from a video.
- We propose a simple NLVL model architecture.
- We establish baselines for the zero-shot NLVL task and compare it with stronger supervision.

2. Related Work

Natural language video localization. Early NLVL works studied relatively constrained environments, such as only cooking events [41]. Recently, large scale, unconstrained NLVL datasets such as Charades-STA [19], ActivityNet-Captions [26] has been appeared. And facilitated by them, there have been advances in deep learning techniques [6, 19, 35, 45, 57], notably in attentive models [39, 43, 44].

However, as the annotations for NLVL are expensive, some literature address weakly-supervised setup of NLVL [12, 21, 33, 38] (WS-NLVL) to alleviate the temporal event annotation. There are various ways to tackle the problem, such as training WS-NLVL as a part of training video captioning [12], building joint visual-semantic embedding framework [38], or selecting among event region proposals [21, 33]. However, although they successfully reduced the temporal annotation cost, the remaining cost of natural language query is still considerable. In contrast, our zero-shot NLVL eliminates both annotations.

Action recognition without annotation. There have been several attempts to classify and localize temporal actions without annotations about actions. Zero-shot action recognition works *et al*. [10, 11, 20, 25, 55] tackled recognizing pre-defined action categories from a video without action labels. Addressing the problem, a number of recent works exploitted object-action co-occurrence patterns from large corpora [10, 11, 25]. These works share similarity with our pseudo-query generation as they utilize co-occurrence patterns of objects and actions. However, our pseudo-supervision generation is more challenging because of the several assumptions they made; they assume the ground truth object labels to be already annotated for the target dataset, action categories to be pre-defined, and videos to be already trimmed according to the ground truth event region [10, 11]. Another line of the works that recognize (or localize) actions without annotated actions is mining action annotations from web [8, 17, 18, 48, 50, 56]. They enable labor-free training of action recognition models [18], but they have potential issues on privacy [14, 51], and often assume weak-level annotations [48, 50] to be exist.

Meanwhile, Soomro *et al*. [47] proposed unsupervised action discovery task to localize actions using only video collections. Jain *et al*. [24] utilized abruptly changing 3D-CNN features to find atomic actions which is combined to compose complex actions. The task that both Soomro *et al*. and Jain *et al*. addresses is similar with our temporal event proposal for generating pseudo-supervisions, but they assume the action classes to be pre-defined and they do not consider relating the actions to language queries.

Grounded language generation. Generating natural language sentences from unlabeled data (*e.g*. sentences with other language, images) addresses similar tasks to our pseudo-query generation. Unsupervised neural machine translation [3, 30, 31] tackled training neural machine translation model without parallel corpora. They partially share the same idea with our pseudo-labeling as they leverage unlabeled sentences for each language to translate. Motivated by the unsupervised neural machine translation (NMT), [15, 29] proposed unsupervised image captioning. Our task setup is partially inspired by this setup; they use an object detector and the independent set of images and sentences to train image captioning, and our zero-shot NLVL uses an off-the-shelf object detector and the independent set of videos and sentences. Compared to these tasks, our task is more challenging since pseudo-supervision includes find-

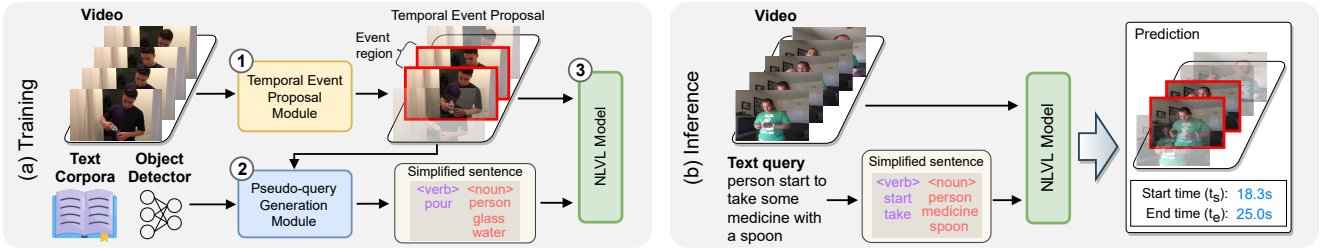

Figure 2: **Overview of PSVL framework for the zero-shot NLVL.** The proposed framework consists of (1) 'temporal event proposal' (TEP), (2) pseudo-query generation (PQ) (3) a supervised NLVL model. (a) At training, pseudo-supervision composed of TEPs and corresponding PQs as a simplified sentence (*i.e.*, nouns and verbs) are generated to train the NLVL model. (b) At inference, a natural sentence query is transformed to a simplified one, and temporal segment boundaries are predicted with the trained model.

ing temporal regions to be queried, in addition to just generating natural language generation. Additionally, we handle videos and sentences, which is more complex than sentences [3, 30, 31] and images [15, 29].

3. Approach

To learn to ground videos to language queries, unlabeled datasets can be utilized in several ways including self-supervised representation learning [49, 60] and generating pseudo-supervisions [15, 32]. The self-supervised learning, however, requires paired supervision of both modalities, but our setup does not provide such paired annotations. Thus, it is not readily applicable to our setup. Instead, we approach this problem by generating pseudo-supervision for training a supervised model, by using text corpora, unlabeled video collections and an off-the-shelf object detector. We name this framework as *Pseudo-Supervised Video Localization* (PSVL) and illustrate it in Fig. 2.

The framework consists of 1) discovering temporal event proposal, *i.e.*, finding event boundaries (Sec. 3.1), 2) generating corresponding pseudo-query (Sec. 3.2), and 3) an NLVL model (Sec. 3.3). One of the benefits of the framework is that any supervised NLVL model such as [39] can be used for this framework. Nevertheless, we further propose a simple NLVL model architecture that is more suitable to the generated pseudo-supervision.

3.1. Temporal Event Proposal

As the first stage of the framework, we discover the temporal event regions of a video that are *meaningful* to be queried. The key challenge here is how to define the notion of *meaningful* temporal segments. We hypothesize that the meaningful events can be selected from a pool of *atomic* temporal regions that can be semantically segmented. Inspired by [24] hypothesizing that frame-wise CNN feature of a video changes abruptly at the event boundaries, we want to discover the *atomic* events, *i.e.*, temporal segments containing a single event. However, the frame-wise features used in [24] only capture the information within that frame

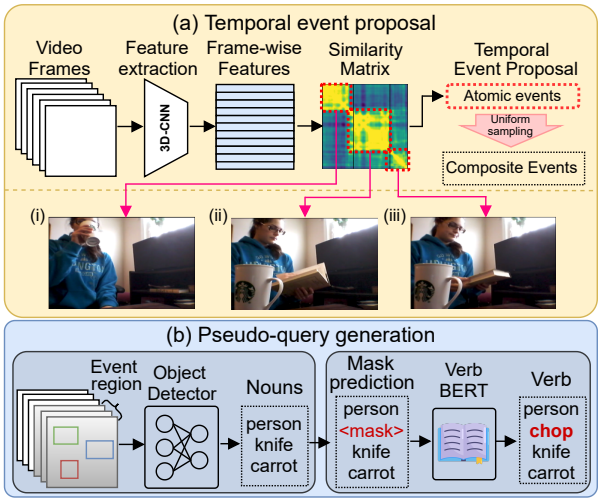

Figure 3: **Generating pseudo-supervision.** (a) event proposal module which uses a self-similarity matrix to cluster and propose event regions. The more the yellow, the more similar to the corresponding frames. (b) pseudo-query generation module.

but miss the contextual information of the video.

To incorporate global context in discovering events, we propose to use a column vector of a similarity matrix of frame-wise visual representation to encode the global information, name as *'contextualized feature'*, similar to [13]. We illustrate the process in Fig. 3-(a). By clustering the contextualized features with frame index using k-means, we generate the atomic events (more details in the supplement).

Meanwhile, the query may require to localize multiple atomic events, *e.g.*, "the person sits then looks at the TV." To address it, we generate a set of *composite* events from the discovered atomic events as the final 'temporal event proposals (TEP).' To generate the composite events, we populate all combinations of *consecutive events*, then sample a few, following a uniform distribution. This simple approach surprisingly results in competitive NLVL accuracy, compared to some recent event proposal methods [24] (Sec. 4.1.1 and more details in the supplement).

Query Type	R@0.3	R@0.5	R@0.7	mIoU
Original Sentence [39]	72.96	59.46	35.48	51.38
Original Sentence (Reprod.)	**73.98**	60.05	**35.75**	**51.63**
Simplified Sentence	73.20	**60.22**	34.30	50.99

Table 1: **NLVL accuracy by different description formats on Charades-STA using LGI model** [39]. 'Reprod.' indicates our reproduction of [39] by authors' implementation[1].

3.2. Pseudo-Query Generation

For each discovered temporal regions (TEP), we generate a corresponding natural language query. We observe that the queries in most supervised datasets are *descriptions* of events of the video segments, *e.g.,* "the person holds doughnut then walks towards door." Generating such descriptive queries can be cast as video captioning [9,26,40,59], involving two challenges; 1) queries should be visually grounded to the temporal region, and 2) queries should be semantically natural. Unfortunately, it requires a large supervised data to train, which is not available in our setup.

Simplified sentence. Instead, we propose to generate *simplified sentence* composed of grounded nouns and inferred verbs, where the nouns are detected by an object detector and the verbs are predicted from language corpora using the grounded nouns. However, as the 'simplified sentence' have only nouns and verbs, is not natural.

In natural language processing (NLP) literature, frame semantic theory [4,16] argues that an event can be expressed as a set of linguistic units such as "*frame elements*" and "*lexical units*", and use them to convey representative semantic meaning of the event. For example, "The person stands up and eats pizza" can be described by "stand eat pizza person." Motivated by this, we relax the problem of generating a natural sentence to generating a set of words, which we call as a 'simplified sentence.'

To empirically validate the effect of the sentence simplification to the NLVL accuracy, we conduct an experiment of converting original query sentences of the supervised NLVL dataset into a simplified one and train a state-of-the-art NLVL model [39] on Charades-STA dataset.

We summarize the results in Table 1; 'Original Sentence (Reprod)' and 'Simplified Sentence' are the performance of the model trained with original (natural) query sentences in the supervised data and corresponding simplified sentence, respectively. We observe that the simplified sentence shows compatible performance to the one with original sentence. This empirically supports that the simplified sentence could be an alternative for describing events for NLVL task.

We now describe how to obtain the grounded nouns and inferred verbs for a video segment in details.

Nouns. Motivated by unsupervised image captioning [15, 29] generating nouns by detecting objects in an image,

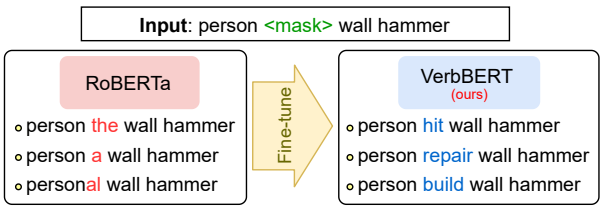

Figure 4: **Predicted verbs by RoBERTa (left) and VerbBERT (right)**. From the contextual words, VerbBERT predicts verbs while RoBERTa [34] predicts any words.

we use an off-the-shelf object detector to obtain nouns grounded to the frames in the temporal region. Note that the off-the-shelf object detector is not trained on the target videos. Therefore, the object detector classes may not include the object of interest in the videos, and the detected objects are often inaccurate; accurate localization but wrong label, or false localization with a random label when the object class is not present in the training dataset of the detector. For reliable object discovery, we only use top-N frequently detected object nouns with high confidence. We investigate the quality of the generated nouns in the supplementary.

Verbs. For predicting verbs, we first consider using pre-trained action recognition model, similar to the object detector for nouns. However, the action labels in action recognition models are much less complete to cover various actionable events in general videos.

As an alternative, motivated by the zero-shot action localization [25], we assume that actions in a temporal region would be constrained by the contextual objects. In other words, since it is likely that both a video frame and a language description would be commonsensical, linguistic statistics would discover appropriate verbs with the surrounding nouns. For instance, if there are some objects such as '*balls*', '*baseball bats*' and '*persons*', the possible verbs would be narrowed down to '*hitting*', '*running*' and *etc*.

Based on this assumption, we propose to infer possible verbs from contextual objects by learning noun-verb co-occurrence patterns in large text corpora. Although the verbs may be deterministically inferred by the contextual objects, the predicted verbs can make the model to attend on the temporal information while objects can be attended for frame-wise information. Note that the verb generation of our task is more challenging than those of zero-shot action recognition in two aspects; first, they assume a closed set of actions to be recognized, but our problem is an open-set problem. Second, the objects are often inaccurate in our sentence, whereas nouns in the contextual objects of zero-shot action recognition [25] are much less noisy.

To efficiently use the large text corpora in a probabilistic model to infer the noun-verb patterns, we want to use a language model (LM) trained on the provided text corpora. But, for a word location, the generic language model pre-

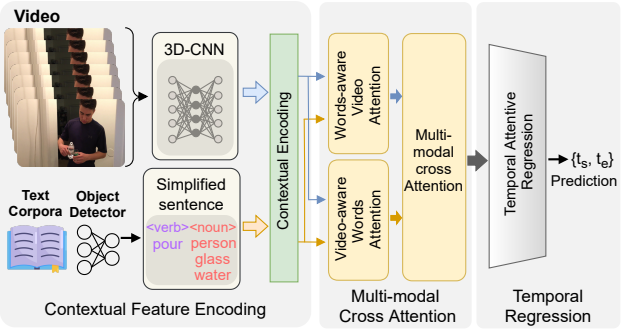

Figure 5: **Overview of the proposed simple NLVL model.** It learned cross modal attentions on the simplified sentence and the proposed temporal event regions to localize events.

dicts not only a verb but also other types of word that is suitable in the context. But we only need a verb for the location. To generate *only* the verbs, we propose to fine-tune a language model (*e.g.*, RoBERTa) to infer verbs only from contextual object nouns, and call it as *VerbBERT*. We describe the details about data collection and fine-tuning procedure for the VerbBERT in the supplement for space sake.

Once VerbBERT is trained, we predict verbs with contextual nouns with a sentence template, following the idea of slot-based captioning methods [28,36,49,54], which provides context words in a fixed template to predict a word at a fixed position (also see supplement for details).

Fig. 4 illustrates a contrasting example of predicting words by RoBERTa and our VerbBERT, showing the advantages of VerbBERT. Given the contextual object nouns such as '*person*', '*wall*', and '*hammer*,' the VerbBERT predicts plausible verbs like '*hit*', '*repair*', and '*build*.'

3.3. A simple NLVL Model

Although any fully supervised NLVL model can be used for our framework, as the pseudo-supervision has less structured sentences, we further propose an NLVL model to be better suited to the simplified sentence input. In particular, we propose a simple attentive cross-modal neural network that learns the sentence structure less but focus more on word-frame attentions as illustrated in Fig. 5. We empirically show that the proposed model slightly outperforms the state of the art NLVL model for fully supervised data, especially in high recall regime (R@0.5 and R@0.7) with less computational cost, in Sec. 4.1.3.

Specifically, the model consists of three parts; 1) contextualized feature encoding to globally encode embedding features of video and simplified sentence input data, 2) multi-modal cross attention network and 3) temporal attentive regression to regress the temporal event region corresponding to the input simplified sentence. For the multi-modal cross attention network, we use query-guided attention dynamic filter [44,57] that fuses the multi-modal information between video and language (Words-aware Video Attention or **WVA**), and a video-guided attention filter to learn the video-aware query embedding (Video-aware Words Attention or **VWA**) followed by a multi-modal cross attention mechanism to fuse all information (Multi-modal Cross Attention or **MCA**). Then, we apply Non-Local block (NL-Block) [52] to encode the global contextual information obtained from the cross-attention module [39]. After the global context features are encoded with each other, we attend on the target temporal segments by temporal attention mechanism [39, 44]. Finally, we predict the temporal boundary regions by a multi-layer perceptron.

Objective function. It consists of two terms; 1) temporal boundary regression loss ($\mathcal{L}_{reg}$) and 2) temporal attention guided loss ($\mathcal{L}_{guide}$) as:

$$\mathcal{L}_{total} = \mathcal{L}_{reg} + \lambda \mathcal{L}_{guide}, \tag{1}$$

where λ is a balancing parameter. Following [39], we use the Huber loss function [23] between the predicted and ground-truth timestamps for $\mathcal{L}_{reg}$ and use a temporal attention guidance loss ($\mathcal{L}_{guide}$) proposed in [39, 44]. More details of the model and the objective are in the supplement.

Inference. As our model is trained with simplified sentence, inference requires translating natural language query into a simplified one. We use an off-the-shelf part-of-speech tagger [22] to convert a sentence to the simplified one.

4. Experiments

Datasets and setups. Following [39, 43], we use two datasets, Charades-STA [19] and ActivityNet-Captions [26], for the NLVL task not using the annotation for training but only for evaluations. We provide a pre-trained Faster R-CNN [42] trained with 1,600 object categories of Visual Genome dataset [27] that is used in [1] as an object detector, and Flicker-description corpus [15] as a language corpus. Further details are in the supplement.

Evaluation metrics. Following [19], we compare the results in two types of metrics: (1) Recall at various intersection over union thresholds (R@tIoU). It measures percentage of predictions that have larger IoU than the thresholds (we use threshold values of $\{0.3, 0.5, 0.7\}$). (2) mean intersection over union (mIoU), which is an averaged temporal IoU between the predicted and the ground-truth region.

Implementation details. Following [39, 43], we extract visual features for each frame using the pre-trained I3D [5] and C3D models [46], respectively. To make the video features fixed-length, we uniformly sample 128 features from a video. For the temporal event proposal, we use $k = 5$ for k-means clustering algorithm for both datasets. We provide a detailed analysis for different k in the supplement.

For generating pseudo-queries, we samples top-5 objects from the object detector and top-3 verbs from VerbBERT to make a simplified pseudo-query. We investigate the effect of different number of nouns and verbs in the supplementary material. To train the VerbBERT, we fine-tune RoBERTa [34] with the Flicker-description corpus.

We match the size of pseudo-supervision data to that of the original supervision, otherwise stated (Sec.4.2). The same-sized supervision makes ours largely comparable to the ones with stronger supervisions. We use $\lambda = 1.0$ as a balancing parameter between losses in Eq.(1).

Baselines. As this is the first work to address zero-shot NLVL, we consider various baselines including 1) predicting random region (**Random**), and ablated methods from PSVL such as a model trained with 2) random query with the proposed temporal event proposal (**Rnd.Q+TEP**), 3) pseudo-query on random temporal regions (**PQ+Rnd.T**), 4) pseudo-query only with the 'grounded nouns' (random verbs) on the TEP (**PQ.N+TEP**) and 5) pseudo-query only with the 'inferred verbs' (random nouns) on the TEP (**PQ.V+TEP**). We use the same NLVL model (Sec. 3.3).

As references, we further present performance of several state-of-the-art weakly-supervised methods such as **TGA** [38], **WSLLN** [21], and **SCN** [33], and fully-supervised methods such as **CTRL** [19] and **LGI** [39]. Note that the weakly-supervised methods [21, 33, 38] are trained with more expensive supervision (aligned pairs of descriptions and temporal regions of a video), whereas ours do not use such paired annotations.

4.1. Quantitative Analysis

We summarize the performance comparison to baselines and methods with stronger supervision in Table 2. In both datasets, clearly the baseline models are much better than the Random method. When the event proposal module is replaced with the temporal event proposal (TEP) and a random query is given, NLVL task performance slightly increases compared to the Random. This implies that without good pseudo-queries, NLVL performance may suffer.

Meanwhile, the PQ+Rnd.T shows high performance compared to the previous two baselines. This implies that although the temporal regions are random, description by PQ allows a model to learn some cross modal representation. When comparing PQ.N+TEP and PQ.V+TEP, we observe that the verb plays a more important role than the noun for the NLVL performance. We believe that this is because the verb contains relationship among contextual objects when describing a temporal region. Our full model (PQ+TEP) outperforms all baselines by significant margins.

Interestingly, PSVL outperforms all weakly-supervised (WS) methods by noticeable margins in Chrades-STA, especially in high recall regime (*e.g.*, R@0.5 and R@0.7) while they outperform ours in mIoU and R@0.3. Similar trend is observed in the experiments on ActivityNet-Captions; ours outperforms all WS methods in R@0.5 (there is no reported results in R@0.7). It implies that our PSVL predicts temporal regions rather precisely while slightly sacrificing overall accuracy (mIoU).

Method	Sup.	R@0.3	R@0.5	R@0.7	mIoU
Charades-STA					
Random	No	26.79	10.82	2.96	17.71
Rnd.Q+TEP		27.39	12.17	1.04	20.12
PQ+Rnd.T		35.31	19.06	6.68	22.95
PQ.N+TEP		28.42	13.18	2.02	24.17
PQ.V+TEP		43.01	20.79	4.97	26.38
PSVL (PQ+TEP)		**46.47**	**31.29**	**14.17**	**31.24**
TGA [38]	Weak	29.68	17.04	6.93	-
WSTG [7]		39.8	27.3	12.9	27.3
SCN [33]		42.96	23.58	9.97	-
CTRL [19]	Full	-	21.42	7.15	-
LGI [39]		72.96	59.46	35.48	51.38
ANet-Captions					
Random	No	23.70	11.41	3.93	16.63
Rnd.Q+TEP		25.98	12.07	4.18	24.12
PQ+Rnd.T		38.19	22.62	7.03	24.92
PQ.N+TEP		30.23	12.92	3.59	25.52
PQ.V+TEP		42.02	23.42	5.91	27.21
PSVL (PQ+TEP)		**44.74**	**30.08**	**14.74**	**29.62**
WS-DEC [12]	Weak	41.98	23.34	-	28.23
WSLLN [21]		42.80	22.70	-	32.20
WSTG [7]		44.30	23.60	-	32.20
SCN [33]		47.23	29.22	-	-
CTRL [19]	Full	28.70	14.00	-	20.54
LGI [39]		58.52	41.51	23.07	41.13

Table 2: **NLVL accuracy on Charades-STA (top) and ActivityNet-Captions (ANet-Captions) (bottom) dataset with various models and supervision level.** 'Sup' refers to supervision level; No (zero-shot), Weak (weakly-supervised [21, 33]), Full (fully supervised). All abbreviations follow the notation in the 'Baseline' paragraph. 'PSVL': our pseudo-supervised video localization method. Among zero-shot methods (No), we highlight the best values in bold and second best in underline.

4.1.1 Temporal Event Proposal

Event proposal methods. We compare PSVL with four baseline temporal event proposal methods in Table 3 (top). ActionByte finds the event boundary utilizing the difference in CNN features between each adjacent frames of video. And, Frame feature uses a method that cluster the similar CNN frame features to generate event proposals.

ActionByte, Frame feature, and our method ('Contextualized feature' by similarity matrix of frame features) outperform random and sliding window by large margins. It implies that both methods discover describable regions for

Event Proposal	R@0.3	R@0.5	R@0.7	mIoU
Random	35.31	19.06	6.68	22.95
Sliding window [33]	35.64	24.84	10.65	24.27
ActionByte [24]	46.55	29.61	12.16	30.06
Frame feature	**48.20**	28.98	11.58	30.76
Contextualized feature (Ours)	46.47	**31.29**	**14.17**	**31.24**
Scoring Function	R@0.3	R@0.5	R@0.7	mIoU
Compactness	45.41	27.82	12.2	29.33
Diversity	**49.41**	22.9	8.71	29.54
Uniform sampling (Ours)	46.47	**31.29**	**14.17**	**31.24**

Table 3: **Temporal event proposal methods.** (top) comparison to other event proposal methods and (bottom) comparison of various scoring functions to aggregate the atomic events to generate candidate (composite) temporal events.

Verb Inference	R@0.3	R@0.5	R@0.7	mIoU
Random verbs	28.42	13.18	2.02	24.17
w/ RoBERTa	34.22	15.49	5.88	25.74
w/ VerbBERT (Ours)	**46.47**	**31.29**	**14.17**	**31.24**

Table 4: **Verb inference methods**. 'Random verbs' are sampled from the verb classes of the VerbBERT model. RoBERTa predicts any words in the missing location, whereas VerbBERT only predicts the verbs by the fine-tuning.

the pseudo queries using visual semantics while others find regions that are either semantically less meaningful or not describable. In addition, we observe particularly large improvements at high threshold recall regime by our method over the others. It implies that the our method finds 'meaningful' events to supervise a model by the help of context.

Scoring functions for composite events. For the atomic event composition (Sec. 3.1), we may use various scoring functions; atomic event's compactness, its diversity, and uniform random sampling to choose top-k composite events. We compare the performance of them in Table 3 (bottom) by various combining function (followed by the PQ generation). Interestingly, the uniform random sampling performs the best. We believe that it contains both compact and diverse combinations of events thus lead to better coverage of training distribution.

4.1.2 Pseudo Query

Effectiveness of VerbBERT. We empirically support the effectiveness of the proposed verb predictor, VerbVERT, by comparing it to RoBERTa [34] and random verbs in Table 4.

For the 'Random verbs' entry, verbs are selected randomly from the set of verbs existing in the large text corpora, and RoBERTa predicts words using the publicly available pre-trained model [53].

As shown in the table, VerbBERT clearly outperforms them as others may generate words other than verbs. It implies that the contextually grounded verbs play a significant role in learning representation related to actions for NLVL.

VWA	WVA	MCA	R@0.3	R@0.5	R@0.7	mIoU
	✓	✓	38.7	21.81	8.24	25.16
✓		✓	42.24	27.91	13.6	28.29
✓	✓		43.15	25.8	12.07	28.67
✓	✓	✓	**46.47**	**31.29**	**14.17**	**31.24**
LGI (Zero-shot) [39]			44.11	28.13	12.87	30.3

Table 5: **Model ablations.** We compare the full model to the ablated ones (word-to-video attention: WVA, video-to-word attention: VWA, multimodal cross attention: MCA), and the current state-of-the-art [39] model trained with the pseudo-supervision.

Number of nouns and verbs. If the number of words is large, it is likely to contain correct signals (high recall), but it may have too much noisy signals from *incorrect* words (low precision) and *vice versa*. We empirically found that five nouns and three verbs results in the best performance. To investigate the trade-off between quantity and quality of words in pseudo-query, we vary the number of objects and verbs. The result is summarized in the supplement.

Quality of generated noun. As mentioned in Sec. 3.2, the nouns from the off-the-shelf object detector is unreliable. To measure how much the noun quality makes the task challenging, we compute average overlaps between the detected objects (η_i) and the original nouns (ξ_i). The recall is 36.54% ($\frac{1}{k}\sum_{i=1}^{k}(\eta_i \cap \xi_i)/\xi_i$, k is the number of descriptions, i is its index). More details are in the supplementary.

Furthermore, we measure the NLVL performance as a function of the number of overlapping objects between detected objects and original descriptions. Specifically, we reduce the overlap ratio by removing the matched nouns and measure the corresponding NLVL performance. As the overlap decreases (36.54% $\rightarrow$ 27.48% $\rightarrow$ 17.97% $\rightarrow$ 9.64% $\rightarrow$ 1.15%), the NLVL 'R@0.5' performance also decreases (31.88 $\rightarrow$ 31.09 $\rightarrow$ 28.25 $\rightarrow$ 25.94 $\rightarrow$ 23.82). This result shows the importance of the overlapping between detected nouns and original description's nouns.

4.1.3 NLVL Model

Ablation on the model components. We investigate the contribution of each component of the proposed simple NLVL model (Sec. 3.3). Specifically, we compare three ablated models by replacing the WVA, VWA, and MCA with simple fully connected layers in Table 5.

Our model outperforms every baseline models by significant margins. Interestingly, the performance drops by ablating the VWA is the largest. Considering the noise in the pseudo queries, we believe that the VWA attention module could suppress noise words in pseudo-queries by visually attending through the VWA attention module. The results of WVA imply that the attention module further filters the noise from the pseudo-queries by attention on the words that are actually meaningful for the given temporal region.

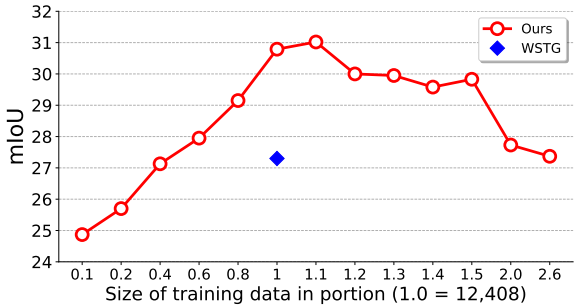

Figure 6: **NLVL performance on Charades-STA on various pseudo-supervision size.** The horizontal axis corresponds to the relative size of pseudo-supervision compared to the original supervision (12,408 samples). Due to the quantity-quality trade-off, the performance peaks at 13,648 samples (1.1x times of the original size). Note that our PSVL even outperforms a recent weakly-supervised model [7] with only 0.6× of the original supervision.

Comparison to the SOTA NLVL model in zero-shot setup. We further compare our model and its ablations with the current state-of-the-art supervised NLVL model [39] with our pseudo-supervision, and call it as 'LGI (Zero-shot).' As shown in the table, our model outperforms [39] in all metrics. We believe this is because [39] is designed to exploit the phrase structure of a natural sentence but the simplified sentence does not have such structure. Moreover, our model is computationally more efficient than the model of [39] for its simplicity. A training iteration of PSVL consumes 0.0664s for 100 samples, whereas the LGI model consumes 0.2s for 100 samples.

4.2. Quantity *vs.* Quality Trade-off

Another benefit of our framework is the ability to generate as many pseudo-supervision data as possible. But the quality of the generated supervision would not be as good as the human supervision. We hypothesize that there exists a trade-off between pseudo-label quality and quantity, similar to the precision-recall trade-off. To empirically verify the trade-off, we conduct an experiment of changing the amount of generated data and compute mIoU on Charades-STA dataset, and summarize the results in Fig. 6.

Until when we provide the data of size of 1.1× of the size of the original supervision data, mIoU monotonically increases, which implies that the quantity prevails quality. Interestingly, with only 60% of the data, ours already outperforms the model trained with weakly supervision. However, when the quantity further increases, the mIOU tends to decreases with a few local increases, implying that the noisy quality of pseudo-supervision prevails the quantity.

4.3. Qualitative Analysis

We present an example of training and inference in Fig. 7. In training (a), PSVL discovers the temporal regions including one with a man wearing a shirt, and produces various nouns and verbs for the region. Among them, some words such as *'man'*, *'shirt'*, and *'wear'* are highly related to the event, but others are not. Our model successfully learns to attend on the words that are correlated to the events, as shown in the words attention weights. At inference (b), the model is able to find a proper temporal region even when a complex query is given. More qualitative results and analyses are available in the supplement.

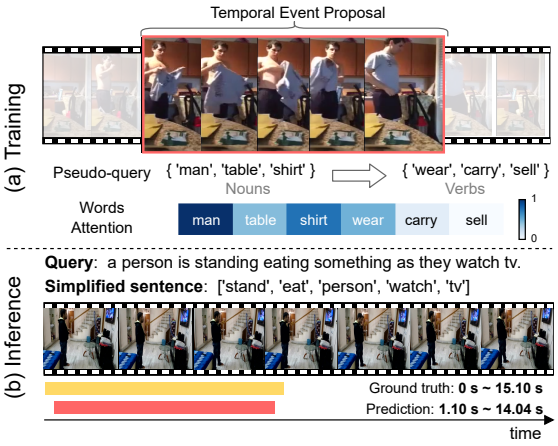

Figure 7: **Qualitative analysis of the training and inference (Charades-STA dataset).** (a) with generated temporal event regions and the visual concepts, the NLVL model is trained to attend meaningful frames and words. We visualize attended weights on words; relatively low in the words *"sell"* and *"carry"* as they are not visually matched. (b) at inference, the NLVL model correctly predicts the temporal boundary with the simplified sentence input.

5. Conclusion

We first propose a novel task of zero-shot natural language video localization. The proposed task setup does not require any paired annotation cost for NLVL task but only requires easily available text corpora, off-the-shelf object detector, and a collection of videos to localize. To address the task, we propose a pseudo-supervised NLVL method, called PSVL, that can generate pseudo-supervision for training an NLVL model. Benchmarked on two widely used NLVL datasets, the proposed PSVL exhibits competitive performance and performs *on par* or outperforms the models trained with stronger supervision.

Acknowledgement. This work was partly supported by NCSOFT, the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.2019R1C1C1009283) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-01842, Artificial Intelligence Graduate School Program (GIST)) and (No.2019-0-01351, Development of Ultra Low-Power Mobile Deep Learning Semiconductor With Compression/Decompression of Activation/Kernel Data, 20%), (No. 2021-0-02068, Artificial Intelligence Innovation Hub) and was conducted by Center for Applied Research in Artificial Intelligence (CARAI) grant funded by DAPA and ADD (UD190031RD).

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018. 5

[2] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, 2017. 1

[3] M. Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. Unsupervised neural machine translation. In *ICLR*, 2018. 1, 2, 3

[4] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The berkeley framenet project. In *the 36th Annual Meeting of the ACL and 17th International Conference on Computational Linguistics*, ACL '98/COLING '98, 1998. 4

[5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 5

[6] Jingyuan Chen, Lin Ma, Xinpeng Chen, Zequn Jie, and Jiebo Luo. Localizing natural language in videos. In *AAAI*, 2019. 2

[7] Zhenfang Chen, Lin Ma, Wenhan Luo, Peng Tang, and Kwan-Yee K. Wong. Look closer to ground better: Weakly-supervised temporal grounding of sentence in video. *ArXiv*, abs/2001.09308, 2020. 6, 8

[8] Nicolas Chesneau, Karteek Alahari, and Cordelia Schmid. Learning from web videos for event classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):3019–3029, 2017. 2

[9] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *CVPR*, 2020. 4

[10] Berkan Demirel, Ramazan Gokberk Cinbis, and Nazli Ikizler-Cinbis. Attributes2classname: A discriminative model for attribute-based unsupervised zero-shot learning. In *ICCV*, 2017. 2

[11] Mandar Dixit, Yunsheng Li, and Nuno Vasconcelos. Semantic fisher scores for task transfer: Using objects to classify scenes. *TPAMI*, 42(12):3102–3118, 2019. 2

[12] Xuguang Duan, Wenbing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang. Weakly supervised dense event captioning in videos. In *NeurIPS*, 2018. 1, 2, 6

[13] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Counting out time: Class agnostic video repetition counting in the wild. In *CVPR*, 2020. 3

[14] Liyue Fan. Practical image obfuscation with provable privacy. In *ICME*, 2019. 2

[15] Y. Feng, L. Ma, Wei Liu, and J. Luo. Unsupervised image captioning. In *CVPR*, 2019. 1, 2, 3, 4, 5

[16] Charles J Fillmore and Collin F Baker. Frame semantics for text understanding. In *NAACL*, 2001. 4

[17] Chuang Gan, Chen Sun, Lixin Duan, and Boqing Gong. Webly-supervised video recognition by mutually voting for relevant web images and web video frames. In *ECCV*, 2016. 2

[18] Chuang Gan, Ting Yao, Kuiyuan Yang, Yi Yang, and Tao Mei. You lead, we exceed: Labor-free video concept learning by jointly exploiting web videos and images. In *CVPR*, 2016. 2

[19] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *ICCV*, 2017. 1, 2, 5, 6

[20] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs. In *AAAI*, 2019. 2

[21] Mingfei Gao, Larry Davis, Richard Socher, and Caiming Xiong. WSLLN:weakly supervised natural language localization networks. In *EMNLP-IJCNLP*, 2019. 1, 2, 6

[22] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python, 2020. 5

[23] Peter J Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, pages 73–101, 1964. 5

[24] M. Jain, A. Ghodrati, and C. G. M. Snoek. Actionbytes: Learning from trimmed videos to localize actions. In *CVPR*, 2020. 2, 3, 7

[25] Mihir Jain, Jan C van Gemert, Thomas Mensink, and Cees GM Snoek. Objects2action: Classifying and localizing actions without any video example. In *ICCV*, 2015. 2, 4

[26] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, 2017. 1, 2, 4, 5

[27] R. Krishna, Yuke Zhu, O. Groth, J. Johnson, Kenji Hata, J. Kravitz, Stephanie Chen, Yannis Kalantidis, L. Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123:32–73, 2016. 5

[28] Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. Babytalk: Understanding and generating simple image descriptions. *TPAMI*, 35(12):2891–2903, 2013. 5

[29] Iro Laina, Christian Rupprecht, and Nassir Navab. Towards unsupervised image captioning with shared multimodal embeddings. In *ICCV*, 2019. 1, 2, 3, 4

[30] Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. In *ICLR*, 2018. 2, 3

[31] Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Phrase-based & neural unsupervised machine translation. In *ICLR*, 2018. 2, 3

[32] Ke Lin, Zhuoxin Gan, and Liwei Wang. Semi-supervised learning for video captioning. In *EMNLP*, 2020. 3

[33] Zhijie Lin, Zhou Zhao, Zhu Zhang, Qi Wang, and Huasheng Liu. Weakly-supervised video moment retrieval via semantic completion network. In *AAAI*, 2020. 1, 2, 6, 7

[34] 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. 4, 6, 7

[35] Chujie Lu, Long Chen, Chilie Tan, Xiaolin Li, and Jun Xiao. DEBUG: A dense bottom-up grounding approach for natural language video localization. In *EMNLP-IJCNLP*, 2019. 1, 2

[36] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *CVPR*, 2018. 5

[37] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*, 2019. 1

[38] Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K Roy-Chowdhury. Weakly supervised video moment retrieval from text queries. In *CVPR*, 2019. 2, 6

[39] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-Global Video-Text Interactions for Temporal Grounding. In *CVPR*, 2020. 1, 2, 3, 4, 5, 6, 7, 8

[40] Jonghwan Mun, Linjie Yang, Zhou Ren, Ning Xu, and Bohyung Han. Streamlined dense video captioning. In *CVPR*, 2019. 4

[41] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *TACL*, 1:25–36, 2013. 2

[42] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 5

[43] Cristian Rodriguez-Opazo, Edison Marrese-Taylor, Basura Fernando, Hongdong Li, and Stephen Gould. Dori: Discovering object relationship for moment localization of a natural-language query in video. In *WACV*, 2021. 1, 2, 5

[44] Cristian Rodríguez-Opazo, Edison Marrese-Taylor, Fatemeh Sadat Saleh, Hongdong Li, and Stephen Gould. Proposal-free temporal moment localization of a natural-language query in video using guided attention. In *WACV*, 2020. 1, 2, 5

[45] Wenbing Huang Peihao Chen Mingkui Tan Runhao Zeng, Haoming Xu and Chuang Gan. Dense regression network for video grounding. In *CVPR*, 2020. 1, 2

[46] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*, 2016. 5

[47] K. Soomro and M. Shah. Unsupervised action discovery and localization in videos. In *ICCV*, 2017. 2

[48] Waqas Sultani and Mubarak Shah. What if we do not have multiple videos of the same action? – video action localization using web images. In *CVPR*, 2016. 2

[49] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid. Videobert: A joint model for video and language representation learning. In *ICCV*, 2019. 3, 5

[50] Chen Sun, Sanketh Shetty, Rahul Sukthankar, and Ram Nevatia. Temporal localization of fine-grained actions in videos by domain transfer from web images. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 371–380, 2015. 2

[51] Ashwini Tonge and Cornelia Caragea. Image privacy prediction using deep features. In *AAAI*, 2016. 2

[52] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *CVPR*, 2018. 5

[53] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 Rush. Transformers: State-of-the-art natural language processing. In *EMNLP*, 2020. 7

[54] Yu Wu, Linchao Zhu, Lu Jiang, and Yi Yang. Decoupled novel object captioner. In *ACM Multimedia*, 2018. 5

[55] Xun Xu, Timothy Hospedales, and Shaogang Gong. Transductive zero-shot action recognition by word-vector embedding. *IJCV*, 123(3):309–333, 2017. 2

[56] Serena Yeung, Vignesh Ramanathan, Olga Russakovsky, Liyue Shen, Greg Mori, and Li Fei-Fei. Learning to learn from noisy web videos. In *CVPR*, 2017. 2

[57] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S. Davis. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *CVPR*, 2019. 1, 2, 5

[58] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, 2018. 1

[59] Luowei Zhou, Yingbo Zhou, Jason J. Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *CVPR*, 2018. 4

[60] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *CVPR*, 2020. 3