Look at What I'm Doing: Self-Supervised Spatial Grounding of Narrations in Instructional Videos

Reuben Tan[1] **Bryan A. Plummer**[1] **Kate Saenko**[1,2] **Hailin Jin**[3] **Bryan Russell**[3]
[1]Boston University, [2]MIT-IBM Watson AI Lab, IBM Research, [3] Adobe Research
{rxtan, bplum, saenko}@bu.edu, {hljin, brussell}@adobe.com
https://cs-people.bu.edu/rxtan/projects/grounding_narrations

Abstract

We introduce the task of spatially localizing narrated interactions in videos. Key to our approach is the ability to learn to spatially localize interactions with self-supervision on a large corpus of videos with accompanying transcribed narrations. To achieve this goal, we propose a multilayer cross-modal attention network that enables effective optimization of a contrastive loss during training. We introduce a divided strategy that alternates between computing inter- and intra-modal attention across the visual and natural language modalities, which allows effective training via directly contrasting the two modalities' representations. We demonstrate the effectiveness of our approach by self-training on the HowTo100M instructional video dataset and evaluating on a newly collected dataset of localized described interactions in the YouCook2 dataset. We show that our approach outperforms alternative baselines, including shallow co-attention and full cross-modal attention. We also apply our approach to grounding phrases in images with weak supervision on Flickr30K and show that stacking multiple attention layers is effective and, when combined with a word-to-region loss, achieves state of the art on recall-at-one and pointing hand accuracies.

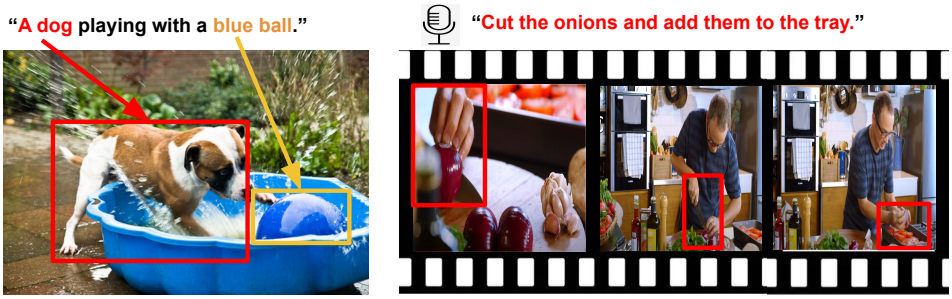
1 Introduction

Content creators often add a voice-over narration to their videos to point out important moments and guide the viewer on where to look. While the timing of the narration in the audio track in the video gives the viewer a rough idea of when the described moment occurs, there is not explicit information on where to look. We, as viewers, naturally infer this information from the narration and often direct our attention to the spatial location of the described moment.

Inspired by this capability, we seek to have a recognition system learn to spatially localize narrated moments without strong supervision. We see a large opportunity for self-supervision as there is an abundance of online narrated video content. In this work, we primarily focus on spatially localizing transcribed narrated interactions in a video, illustrated in Figure 1 (right). Unlike prior phrase grounding work (Figure 1 left), which primarily focuses on matching a noun phrase to an object, our task involves matching entire sentences to regions containing multiple objects and actions.

Not only is this task integral to advancing machine perception of our world where information often comes in different modalities, it also has important implications in fundamental vision-and-language research such as robotics, visual question answering, video captioning, and retrieval. Our task is challenging as we do not know the correspondence between spatial regions in a video and words in the transcribed narration. Moreover, there is a loose temporal alignment of the narration with the described moment and not all words refer to spatial regions (and vice versa). Finally, narrated interactions may be complex with long-range dependencies and multiple actions. In Figure 1 (right), notice how "them" refers to the onions and that there are two actions – "cut" and "add".

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.



Phrase Grounding (prior work) Narrated Interaction Localization (this paper)

Figure 1: Prior work on *phrase grounding* focuses on localizing single objects in static images (left, example from Flickr30K [35]). In this paper, we propose a new task of spatially *localizing narrated interactions* in video (right). In the task of phrase grounding, the goal is to localize nouns or adjective-noun phrases in images. In contrast, interaction grounding may involve multiple objects and actions and use entire sentences. Our approach self-trains on a large corpus of narrated videos and does not require supervision. (Video credit: Woolworths [50])

Contrastive learning [33] and computing contextual features via an attention model [46] are natural tools for addressing the aforementioned alignment and long-range cross-modal dependency challenges. While contrastive learning has been successfully applied for temporal alignment of a video clip with a narration [30], the alignment of spatial regions with narrations has not been addressed. Naively applying a joint attention model over the set of features for the two modalities, followed by optimizing a contrastive loss over the two modalities' aggregated contextual features, results in a fundamental difficulty. Recall that an attention computation comprises two main steps: (i) selecting a set of features given a query feature, and (ii) aggregating the selected features via weighted averaging. By selecting and aggregating features over the two modalities during the attention computation and subsequently optimizing a contrastive loss, the model in theory may select and aggregate features from only one modality to trivially align the contextual representations.

While there are approaches for grounding natural language phrases in images without spatial supervision, the best-performing approaches involve matching individual words to regions [19] or not using any paired text-image supervision at all [48]. This strategy is effective for simpler short phrases, as illustrated in Figure 1 (left). However, consider the interaction "cut the onions and add them to the tray", shown in Figure 1 (right). This interaction is described with more complex, compositional language. In the first two frames, there are multiple "onion" objects visible and the correct one depends on the action being applied. Unlike the task of object grounding, it is not sufficient to simply co-localize all of the mentioned objects. We will show that it is difficult to effectively employ a strategy that matches individual words to regions to localize such complex interactions.

To address these challenges, we make the following contributions. First, we learn to localize the spatial location of a narrated interaction from abundantly available narrated instructional videos [31]. Our approach does not require manually collecting sentence descriptions or the locations of the described interactions for training. Here, the automatically transcribed narrations allow for self-supervised learning via alignment with the video clip.

Second, we propose a new approach for directly contrasting aggregated representations from the two modalities while computing joint attention over spatial regions in a video and words in a narration. We show that optimizing a loss over the aggregated, sentence-level representation allows for a global alignment of the described interaction with the video clip, and offers an improvement over optimizing a matching loss over words and regions. To overcome the network learning a trivial solution while directly contrasting jointly attended features from the two modalities during training, we design network attention layers that do not allow feature aggregation across the two modalities. Our strategy involves alternating network layers that compute inter- and intra-modal attention. Our inter-modal attention layer allows features from one modality to select and aggregate features from only the other modality, and not within its own modality. Our intra-modal attention layer selects and aggregates features from within the same modality. This strategy ensures that the output representations do not aggregate features across the two modalities. In combination with stacking the inter- and intra-modal attention layers, our approach attends jointly and deeply and directly contrasts the resulting contextual representations from the two modalities.

Finally, we introduce an evaluation dataset that provides bounding box annotations for interactions described by natural language sentences. Our dataset is built on the validation split of the YouCook2 dataset [58] and contains approximately 1000 segments of varying durations. We demonstrate our approach on our collected evaluation dataset of localized interactions and on localizing objects via the YouCook2-BB benchmark where we show that we outperform shallow and full cross-modal attention. We also apply our approach to grounding phrases in images with weak supervision on Flickr30K and show that stacking multiple attention layers is effective with our loss. Furthermore, we show that our loss is complementary with the word-to-region loss of Gupta *et al*. [19], and when combined with it, achieves state of the art on recall-at-one and pointing-hand accuracies.

2 Related Work

Self-supervised learning has been a significant factor in unprecedented progress in natural language processing, resulting in effective and robust learned word representations [32, 14, 53, 34] that achieve state-of-the-art performance on downstream tasks. Recently, it has also garnered a lot of interest in the computer vision community, achieving state-of-the-art performance in unsupervised pre-training of deep visual models. [9, 21, 10] have demonstrated that effective image representations can be learnt simply by contrasting between augmented views of the same image without labels. [55, 40] found that the resulting representations are biased by background pixels and tried to improve foreground object localization via data augmentation or saliency. In video, self-supervised tasks like pace prediction ad future prediction [47, 20, 16, 23, 6, 49] were used for pre-training, while [2] used self-supervision on audio-visual data for tasks like unsupervised speaker localization. In this work, we propose a self-supervised method to spatially localize activity descriptions in video.

Object and action localization Approaches generally leverage a region proposal network (RPN) [45, 18, 17] as well as region-based convolutional neural network (CNNs) to detect objects or localize actions temporally [52] or spatio-temporally [24]. However, they are often trained on curated datasets and are, consequently, limited by a fixed predefined number of object categories in these datasets. [58] collected a dataset of cooking videos labeled with descriptions and temporal segments but did not address spatial grounding of activities. Motivated by this limitation, our proposed approach aims to learn to recognize human interactions with objects that belong to the long-tailed distribution from uncurated and unlabeled online videos. There have been video datasets collected with bounding box annotations with associated natural language descriptions [56]. These datasets provide bounding boxes for noun phrases [12, 56] or for egocentric videos [12]. We collect the first dataset of captioned interactions annotated with spatial bounding boxes for evaluating narrated interactions.

Vision-and-language. Semantic information from natural language has often been exploited to provide an additional source of supervision for learning visual representations. Such approaches have generally leveraged image [26, 28, 35] and video [5] datasets that are annotated with natural language descriptions to learn a joint embedding space, where visual and language representations of semantically-similar pairs are close. Prior approaches have mainly focused on retrieving short clips from a large video corpus [30] or localizing relevant segments within untrimmed videos [5, 8, 11]. However, they do not identify the relevant spatial locations of the described interactions. Some existing works aim to spatially localize a natural language query [39, 54, 27] but focus on objects and train on ground truth bounding boxes, whereas our approach is unsupervised. Object grounding with weak or no supervision is addressed in [11, 4, 7]. To alleviate the costly annotations of curated datasets, [30] have proposed to learn robust video representations as well as an effective joint video-text embedding space from a large corpus of unlabeled and uncurated instructional videos [31]. In concurrent work, CLIP [36] trained image-text representations on a dataset of 400 million (image, text) pairs collected from the internet. These efforts to learn from "free" data are similar in spirit to our work, however, we focus on the localization task.

3 Self-Supervised Grounding of Narrated Video Interactions

Given a video clip and a narration, our objective is to learn to localize the relevant spatial regions in the clip that correspond to the interaction described in the narration without bounding box supervision. To this end, we propose a novel Contrastive Multilayer Multimodal Attention (CoMMA) module (Figure 2(c)) that allows for direct early interactions between the spatiotemporal regions and words,

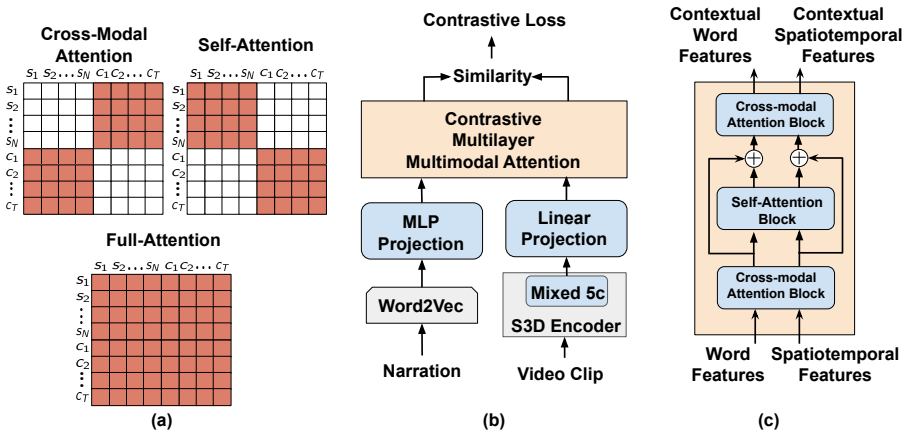


Figure 2: **Approach overview.** (a) Binary masks for computing the different types of cross-modal attention. The shaded regions represent "true" Boolean values where a query feature may select a key. If the mask value is "false", then that corresponding (query, key) pair is not considered in the computation of attention and weighted feature aggregation. (b) Our proposed architecture for computing joint attention over narrations and videos. (c) Our proposed **Contrastive Multilayer Multimodal Attention (CoMMA)** module.

as well as an effective contrasting of the two modalities' signals. We define two modalities to have early interactions when their features attend to each other to compute their final contextualized representations. Conversely, in a late alignment mechanism, the two modalities only interact when the training loss is computed. The intuition underlying our proposed approach is that a multimodal attention model with multiple stacked layers will encourage a more robust and fine-grained alignment between weakly aligned video clips and sentences. Through repeated early interactions of the two modalities and aggregation of unimodal contextual information, it allows directly contrasting the final representations for the two modalities and becomes more discriminative of irrelevant regions and words. Our approach goes beyond shallow attention architectures [3] and late fusion of the two modalities [30]. Our multimodal attention module can be easily stacked on top of a base convolutional feature encoder for localization.

Let c_t be a spatiotemporal feature for region t and s_i be a word feature for the i-th word in a narration with N words. We obtain features c_t and s_i via base features that are projected to a common embedding space using linear and MLP layers. We form matrix inputs C_0 and S_0 by stacking the spatiotemporal and word features c_t and s_i, respectively, as column vectors.

3.1 Contrastive Multilayer Multimodal Attention

The deep attention module comprises alternating layers of a bidirectional cross-modal attention layer followed by a self-attention layer (Figure 2(c)). This architecture choice is based on the reasoning that a hierarchical early interaction process may encourage a model to learn the fine-grained alignment between video regions and narrations while ignoring the irrelevant parts. More specifically, each bidirectional cross-modal layer computes new representations for a target modality with the latent representations from the source modality to learn the relevance of each region to each word or phrase, and vice versa. Additionally, the self-attention layer serves to identify relevant words and regions by aggregating contextual information between the augmented unimodal features.

The bidirectional cross-modal attention layer can be formulated as the key/query/value attention mechanism that is commonly used in attention models [46]. The cross-modal setting provides a natural choice for this attention mechanism where the queries are spatiotemporal region features and the keys and values are word features, and vice versa. The attention layer accepts a concatenated sequence of word and region features and a set of binary attention masks. In Figure 2(a), the shaded regions of the masks provide an indication for the attention layer to consider the corresponding elements in the query and key in its computation. Specifically, the cross-modal attention (CA) mask M_{CA} computes the attention over the words with respect to each region and vice versa. Note that this differs from the full-attention (FA) mask M_{FA} (illustrated in Figure 2(a) bottom) that computes attention over the *entire* multimodal sequence of words and regions. We adopt the attention

mechanism in [46] where the attention weights are first computed by measuring the similarity between the query and key before computing a weighted sum of the value representations. Formally, let Q, K, V be the query, key and value matrices, respectively. Let the masked attention mechanism be represented as:

$$Attn(K, Q, V, M) = V \operatorname{softmax}\left(\frac{(Q^\top K) \odot M}{\sqrt{D}}\right) \tag{1}$$

where M is any one of the binary masks shown in Figure 2(a), D denotes the dimensionality of the query, and $\odot$ is the Hadamard product.

For simplicity, let matrix $Y_0^{SA} = [C_0, S_0]$ denote the stacked input features from both modalities. In addition, let $W_K^{i_l}$, $W_Q^{i_l}$ and $W_V^{i_l}$ denote the projection matrices for keys, queries and values, respectively, for attention type i and layer l. Here, i can be used to represent a cross-attention (CA) or self-attention (SA) layer and l is the layer number. We compute the output of the first cross-modal attention layer by plugging in the projected video and narration features:

$$Y_{l+1}^{CA} = CrossAttn(Y_l^{SA}) = Attn(W_K^{CA_{l+1}} Y_l^{SA}, W_Q^{CA_{l+1}} Y_l^{SA}, W_V^{CA_{l+1}} Y_l^{SA}, M_{CA}) \tag{2}$$

where l indicates the layer number and M_{CA} is the cross-modal attention mask. Note that if input to Equation 2 is $Y_l^{SA} = [C_l, S_l]$, then the output will be $Y_{l+1}^{CA} = [S_{l+1}, C_{l+1}]$, *i.e.*, (contextualized) video features become (contextualized) language features. The similarity scores computed between the queries (regions) and the keys (words), depicted as the normalized dot product between the query and key in Equation (1), act as a soft attention mechanism to measure the relevance of each word with respect to a region. The softmax-normalized scores are multiplied with the value (words) vectors to compute a region-specific narration representation. To illustrate this concept, consider a YouTube video clip about cooking fried chicken. The words 'fried' and 'cook' will have a higher similarity score than 'subscribe' since some regions will be highly relevant to the former. Conversely, the modalities of the keys and queries can be swapped to compute word-specific clip representations.

We emphasize that in the cross-attention layer, there is no mixing of features from different modalities where an elementwise sum is applied to features across the different modalities. Consequently, it allows both modalities to attend to each other without leaking any information between them. We include a mathematical formalization of this operation in the supplementary.

To aggregate contextual information over the augmented representations of the two modalities, the outputs of the cross-modal attention layer are passed into a self-attention layer. The self-attention mask M_{SA} is used to ensure that attention is only computed between elements of the same modality. Formally, this computation can be represented as

$$Y_l^{SA} = SelfAttn(Y_l^{CA}) = Attn(W_K^{SA_l} Y_l^{CA}, W_Q^{SA_l} Y_l^{CA}, W_V^{SA_l} Y_l^{CA}, M_{SA}) + Y_l^{CA} \tag{3}$$

where M_{SA} is the self-attention mask. Its output is passed into the next cross-attention layer. Let C_L and S_L denote the outputs of the last layer in the proposed cross-modal attention module after L layers where $C_L \in \mathbb{R}^{DXT}$ is the set of contextualized spatiotemporal region features and $S_L \in \mathbb{R}^{DXN}$ is the set of contextualized word representations.

Comparison to multimodal attention models. To evaluate the importance of multimodal attention modules for localizing narrations, we briefly compare CoMMA against other state-of-the-art variants. [3] is an image-level approach that utilizes features from different levels of an encoder. However, unlike CoMMA, it only utilizes a single round of cross-modal interaction. Another highly relevant work is the Contrastive Bidirectional Transformer (CBT) [44]. Despite the similarities in attention layers, the CBT model computes self-attention over the entire multimodal sequence (similar to the full-attention mask in Figure 2(a)) and aggregates a summary knowledge into a sentinel vector. In contrast, CoMMA allows for repeated early interactions without mixing of features from different modalities. We show in Section 4 that this is critical for localizing from grid features.

3.2 Contrastive loss and inference

We train our proposed multimodal attention module by contrasting between positive and negative pairs of video clips and narrations. Formally, we aim to learn language-aligned visual representations for video clips such that features for corresponding videos and narrations are similar and non-corresponding features are dissimilar. The final clip representation $\hat{C}$ and sentence representation

$\hat{S}$ are computed by mean-pooling over the spatiotemporal regions $\hat{C} = \frac{1}{T}\sum_{t=1}^{T} C_{L,t}$ and words $\hat{S} = \frac{1}{N}\sum_{j=1}^{N} S_{L,j}$ respectively. The final contrastive objective is optimized:

$$\mathcal{L}_{sentence} = -\sum_{i=1}^{n} \log\left(\frac{\exp((\hat{C}_i)^T \hat{S}_i)}{\exp((\hat{C}_i)^T \hat{S}_i) + \sum_{\hat{S}_{neg}\sim\mathcal{N}_{i,S}} \exp((\hat{C}_i)^T \hat{S}_{neg}) + \sum_{\hat{C}_{neg}\sim\mathcal{N}_{i,C}} \exp((\hat{C}_{neg})^T \hat{S}_i)}\right) \quad (4)$$

where the negative sets $\mathcal{N}_{i,C}$ and $\mathcal{N}_{i,S}$ comprises non-corresponding video clip and narration pairs for the j-th sample, and n denotes the total number of training samples. Note that cross-modal attention is applied to negative pairs as well. In contrast to the word-level objective (defined in Section 4) used in [19] which contrasts word features separately, contrasting aggregated features in the sentence loss is more robust to the weak temporal alignment between narrations and videos.

Inference. During inference, given a natural language sentence and a video clip, we aim to localize the salient spatial regions across the frames in the clip. Beginning from the attention weights of the last cross-modal attention layer, we apply attention rollout [1] to obtain the final attention heatmap over all spatiotemporal regions. In attention rollout, the attention weight matrices from all cross-attention and self-attention layers are multiplied recursively to yield the output localization scores A for each spatiotemporal region, where $A = \prod_{l=0}^{L} W_l$ where A and W_l are the final attention heatmap and attention weights from the l-th layer, respectively.

The resulting localization scores aggregate the total amount of attention by the entire set of spatiotemporal and word features assigned to a query feature. Since the temporal and spatial dimensions of the attention heatmap have been downsampled from the original input resolution, the localization scores are temporally and spatially interpolated back to the input resolution. We return the final spatiotemporal location as the mode over all output scores.

4 Experiments

We evaluate and ablate our method on two video tasks – spatial localization of interactions and objects – and weakly supervised phrase grounding in images.

Word-level loss. We experiment with a word-level loss as used in state-of-the-art phrase grounding [19], which aims to learn an alignment between each word in the narration and all spatial regions. As shown in [19], the word-level loss has been shown to be effective at localizing noun phrases in images. Intuitively, it seeks to maximize the mutual information between each word and the region features if the narration corresponds to the clip. Specifically, the word-level loss is formulated as:

$$\mathcal{L}_{word} = -\sum_{i=1}^{n}\sum_{j=1}^{N} \log\left(\frac{\exp\left((S_{i,L,j})^T S_{i,j}^{val}\right)}{\exp\left((S_{i,L,j})^T S_{i,j}^{val}\right) + \sum_{m\sim\mathcal{N}_i} \exp\left((S_{m,L,j})^T S_{i,j}^{val}\right)}\right) \quad (5)$$

where the negative set $\mathcal{N}_i$ of the i-th sample comprises the word features that are attended to by non-corresponding video clips and N denotes the number of words in the narration, and n denotes the total number of samples in a training batch. $S_{i,L,j}$ denotes the representation output of our CoMMA for the j-th word of the i-th sample. Finally, for a given input word $S_{i,0,j}$, we compute its value representation as follows: $S_{i,j}^{val} = MLP(S_{i,0,j})$.

4.1 Grounding interactions and objects in video

Self-training dataset. We self-train our proposed model on instructional videos from the HowTo100M dataset [31]. This dataset contains approximately 1.2 million instructional videos that are sourced from YouTube, encompassing multiple domains such as cooking and hand-crafting. These videos contain narrations that are either uploaded manually by users or the output of an Automatic Speech Recognition (ASR) system. Furthermore, the narrations have been pre-processed to remove all stop words. As noted in prior work [30], there is a weak temporal alignment between

the narrations and the video clips, adding another degree of difficulty in training with such videos. The misalignments are due to the nature of instructional videos where an interaction is often narrated before or after demonstrating it. To reduce the computational burden of training over the entire 100M clips, we identify a set of clips from the dataset that roughly aligns with the YouCook2 dataset. To achieve this goal, we extract a list of verb and nouns from the vocabulary of YouCook2 and filter HowTo100M video clips with narrations that contain these words. Our final pre-training set comprises approximately 250,000 video clips. Note that our base video and language features were trained on the entire HowTo100M dataset, which provides a good initialization for our learning task.

Evaluation datasets. We introduce a new evaluation dataset, YouCook2-Interactions, that provides spatial bounding box annotations for interactions that are described by a natural language sentence. Our dataset is built on the validation split of the YouCook2 dataset [58]. We describe our dataset in full in the supplemental. We evaluate using our collected YouCook2-Interactions dataset for localizing described interactions.

In addition to localizing interactions given a sentence description, we also consider localizing objects given a single object keyword. For localizing single objects, we evaluate on the YouCook2-BB dataset [57]. The YouCook2-BB dataset is constructed by augmenting the original YouCook2 dataset with bounding box annotations for object locations. In our experiments on YouCook2-BB, we only evaluate our approach on the validation split since the test split annotations are not readily available.

Evaluation criteria. For both video tasks, we evaluate the quality of the output detections using the pointing hand accuracy. Specifically, given a sentence and the ground-truth bounding box, we consider our model to have a "hit" if the pixel with the highest co-attention similarity score lies within the ground-truth bounding box. Otherwise, it is a "miss". The final localization accuracy is computed as follows: Localization Accuracy $= \frac{\text{\# hits}}{\text{\# hits}+\text{\# misses}}$.

Implementation details. We use the Separable 3D CNN-Gated (S3D-G) encoder [51] as well as a shallow language encoder for our video and language models, respectively. To train our proposed model, we set a learning rate of 1e-4 and optimize the model using the AdamW [29] optimizer. An epoch equivalent of steps is used for linear warmup. In our experiments, we also explore using 2 CoMMA modules but due to the large memory requirements, the batch size has to be reduced significantly. The results from our initial experiments and prior work on contrastive learning show that a large batch size is crucial to achieving strong performance. More details are in the supplemental.

Baselines. In our evaluations, we compare our approach against baselines such as a center prior and optical flow, as well as state-of-the-art image-grounding models. In the center prior, we return the center pixel for each frame. The center prior is a simple yet competitive baseline that performs well due to the nature of these videos, where the main subject is generally in the center of the camera view. In addition, we also implement an optical flow baseline. An optical flow map is generated for each frame using the OpenCV library and the magnitudes are averaged across both spatial directions. Finally, we select the pixel with highest average magnitude as the location prediction.

In addition, we compare against four cross-modal attention baselines. First, we adapt the pretrained MILNCE [30] model to remove the penultimate global average pooling layer in the S3D-G video encoder. Consequently, the output of the video encoder is a set of spatiotemporal features that have been projected into the joint embedding space. Then, we compute the dot product between the sentence feature and the spatiotemporal features before performing linear interpolation to determine the mode pixel. Second, we compare against a baseline that leverages full attention with a sentinel vector ("Full attention + sentinel") that predicts the similarity of the two modalities' features, which is then passed to the InfoNCE loss (inspired by CBT [43]). Third, we compare our multilayer model against a shallow cross-modal attention layer ("Shallow attention") inspired by Akbari *et al*. [3]. To maintain fair comparisons, we implement a similar version of their model except we use for the base features the convolutional feature map outputs of the S3D video encoder. Finally, we include comparisons against a state of art phrase grounding model [19]. We apply the model out of the box to our task with the exception of using our features for fair comparison.

Interaction localization. We report in Table 1 results of our approach and the baselines on the interaction localization task. Our proposed approach outperforms the center prior baseline by a significant margin of 20% on the task of localizing interactions given natural language sentences. Surprisingly, computing the region with the maximum degree of motion through optical flow does not

Table 1: **Interaction localization evaluation on our YouCook2-Interactions dataset.** Our approach outperforms baselines, including shallow and full attention.

Approach	Training Loss	Localization Accuracy (%)
Center Prior	-	31.81
Optical flow	-	18.75
MILNCE [30]	Sentence level	27.18
Contrastive phrase grounding [19]	Word level	24.04
Shallow Attention	Sentence level	48.30
Full attention + sentinel	Sentence level	39.03
CoMMA (Ours)	Word level	34.91
CoMMA (Ours)	Sentence level	**55.80**

Table 2: **Spatiotemporal self-attention ablation and attention layer ablation.** (Left) We ablate different types of self-attention over spatiotemporal features and find computing full self-attention over spatiotemporal features works best. (Right) We report an ablation of the different attention module components in our model. We find that multiple stacked layers outperform shallower architectures.

Self-Attention Type	Localization Accuracy (%)
Spatial	53.88
Temporal	52.72
Spatial + Temp.	52.15
Spatiotemporal	**55.80**

Cross-Modal Attention 1	Cross-Modal Attention 2	Self-Attention	Localization Accuracy (%)
✓	✗	✗	48.30
✓	✓	✗	51.09
✓	✓	✓	**55.80**

provide a good proxy for localization. The large performance gain achieved by our proposed approach over the optical flow baseline suggests that tracking regions with maximum movement is insufficient for video understanding. In addition, we observe that our proposed approach outperforms a state-of-the-art phrase grounding model [19] by a large margin. One possible reason is that enforcing an alignment between words and regions is not suitable for grounding interactions in videos, especially when there is a weak temporal alignment between clips and narrations. This is corroborated by a huge drop of 20% in localization accuracy when our CoMMA is trained with the word-level loss instead.

The difference in performance obtained by our model and the pretrained MILNCE model suggests that a model trained for retrieval may not be focusing on the relevant regions as described by the natural language sentence. As seen in Table 1, our proposed approach outperforms the "full attention + sentinel" baseline by 15%. This result suggests that aggregating contextual information into a sentinel vector causes the model to lose fine-grained information required for localization. We note that, simila r to [43], the CBT baseline network outputs a "Mutual Information (MI)-like score" that is plugged in directly to the NCE loss; it does not compute a cosine similarity score between the video clip and natural language query features. We hypothesize that computing this MI-like score (instead of directly contrasting the two modalities' features) may somewhat mitigate the CBT baseline from catastrophic failure. We suspect that this is due its cross-modal attention mechanism where features from both modalities are mixed during the interaction process.

Multilayer attention and spatiotemporal self-attention ablations. We report in Table 2 (Left) an ablation over stacking the different attention layers in our network. Notice in rows 1 and 2 that the localization accuracy increases as we stack more attention layers in our network, suggesting that deeper cross-modal attention models help to improve the latent alignment between different modalities. Results from an ablation of the different forms of self-attention over spatiotemporal regions are reported in the right table. It is interesting to observe that by applying self-attention to spatial regions alone, the model performs competitively by obtaining a localization accuracy of 53.88%. We obtain our best model when self-attention is computed over all spatiotemporal regions. This suggests that reasoning about temporal context is important but it has to be done in conjunction with spatial context.

Object localization. Besides localizing full narrated sentences, we also evaluate our model's capability to localize single objects. Specifically, we compare it against the state-of-the-art Not All Frames Are Equal (NAFAE) model [42]. We report results in Table 3. Despite being trained with

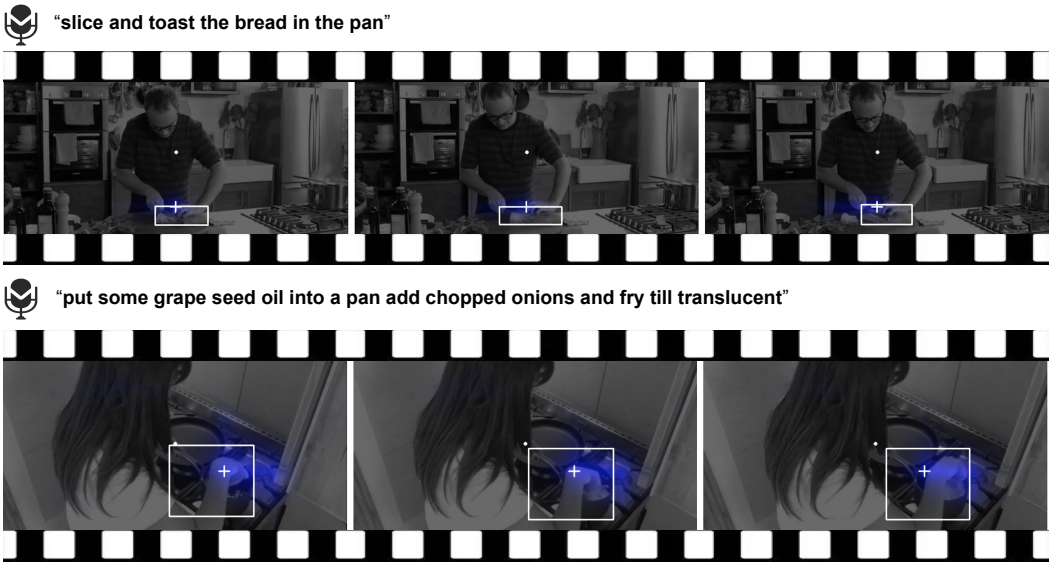


Figure 3: Examples of correct localization predictions by our model. Our approach localizes narrated interactions involving a person's hands and interacted objects. (Video credit: Woolworths [50] and Alicia Kirby [25])

sentence-level loss, we observe in Table 3 that CoMMA localizes objects corresponding to single noun words. Notably, our proposed approach outperforms NAFAE. We give more details of the NAFAE baseline in the supplemental.

Qualitative results. We show qualitative results in Figure 3. Each frame is overlaid with the computed attention heatmap to depict the visual cues that our model uses to make its prediction. The white circle indicates the location of the center prior and the cross denotes the mode pixel of the computed attention heatmap. Notice how our approach correctly localizes the described interaction.

Table 3: **Object localization evaluation on the YouCook2-BB dataset.** By self-training on narrations, our approach learns to localize objects and outperforms the baseline.

Approach	Full Loc. (%)
NAFAE [42]	46.95
Ours	**59.25**

4.2 Weakly supervised phrase grounding in images

We evaluate our approach on the task of weakly supervised phrase grounding to determine its capability to generalize to localizing noun phrases in images. For this task, we adopt the same setup as in prior work [19, 13, 15] by training on MSCOCO [28] and evaluating on Flickr30k [35].

We report results on the standard recall-at-K (R@K) and pointing accuracy criteria [19]. For training, we optimize the sum of the word-to-region matching loss and our sentence loss: $L(\theta) = L_{word} + \lambda L_{sentence}$. We set λ to be 0.005 in our experiments. We compare our approach against state-of-the-art phrase grounding models. Details of how the model in [19] is extended with our CoMMA architecture as well as a description of the baselines are inclided in the supplemental. We note that the the value projection layer for region features in [19] is omitted to incorporate CoMMA. Finally, we build on the same set of visual and language features from [19] for fair comparison.

Results and analysis. We report results in Table 4. As mentioned above, omitting the region value projection layer results in a performance drop as evidenced by the results obtained by using a single CA layer. Adding a second cross-attention layer generally helps to improve localization accuracy. It leads to notable increases in R@1 and pointing-recall accuracies, surpassing state-of-the-art [19] on those criteria. We note that our full CoMMA model with unimodal self-attention layers hurts performance (R@1 = 48.81%). One possible reason is that optimizing CoMMA on this task is challenging and it may be overfitting. However, it is very promising that adding more layers works out of the box in general on a task that is typically addressed by shallow attention models. In contrast,

Table 4: **Weakly supervised phrase grounding experiments on Flickr30K.** Adding additional cross-attention layers improves performance on the R@1 and pointing recall criteria. We include a description of the visual features and baseline approaches in the supplemental. IN and VG denote ImageNet and Visual Genome respectively.

Approach	Visual Features	R@1	R@5	R@10	Pt Recall
Fang *et al* [15]	VGG-cls (IN)	-	-	-	29.00
Akbari *et al* [3]	VGG-cls (IN)	-	-	-	61.66
Akbari *et al* [3]	PNAS Net (IN)	-	-	-	69.19
Align2Ground [13]	Faster-RCNN (VG)	-	-	-	71.00
Gupta *et al* [19]	Faster-RCNN (VG)	51.67	**77.69**	83.25	76.74
ours (word + sent): 1 CA layer	Faster-RCNN (VG)	49.99	76.72	**83.51**	74.97
ours (word + sent): 2 CA layers	Faster-RCNN (VG)	**53.80**	76.69	82.28	**76.78**

we show in the supplemental that stacking multiple layers in our default CoMMA architecture when trained with the sentence-level loss $L_{sentence}$ only results in improved R@K localization accuracies.

Finally, we also observe that, unlike the observations in the video experiments, the word-level loss is critical for phrase grounding in images. We hypothesize this is because the captions are strongly correlated with the images and consist primarily of noun phrases. In contrast, the video task requires spatially localizing different interactions that are mentioned in a narration, which are more semantically complex as compared to noun phrases, making the word loss less effective.

Discussion. This work introduces the problem of localizing narrated interactions in uncurated videos without bounding box supervision or priors. To address this, we propose a novel Contrastive Multi-layer Multimodal Attention module that facilitates repeated early interactions between spatiotemporal regions and words to learn their latent alignment. Additionally, we introduce a new evaluation dataset that provides spatial annotations for narrated instructions in videos. Evaluations on the three separate tasks of interaction and object localization in videos as well as phrase grounding in images demonstrate the ability of our model to ground both words and phrases more effectively than the shallow and full-attention baselines. Our approach opens up the possibility of further exploration in other modalities. Learning to associate relevant spatial regions with natural language sentences may also be beneficial to learning richer video representations, particularly with the recent gravitation towards self-supervised representation learning.

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020.

[2] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. *ECCV*, 2020.

[3] Hassan Akbari, Svebor Karaman, Surabhi Bhargava, Brian Chen, Carl Vondrick, and Shih-Fu Chang. Multi-level multimodal common semantic space for image-phrase grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12476–12486, 2019.

[4] Elad Amrani, Rami Ben-Ari, Inbar Shapira, Tal Hakim, and Alex Bronstein. Self-supervised object detection and retrieval using unlabeled videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 954–955, 2020.

[5] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017.

[6] Sagie Benaim, Ariel Ephrat, Oran Lang, Inbar Mosseri, William T Freeman, Michael Rubinstein, Michal Irani, and Tali Dekel. Speednet: Learning the speediness in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9922–9931, 2020.

[7] Gedas Bertasius and Lorenzo Torresani. Cobe: Contextualized object embeddings from narrated instructional video. *arXiv preprint arXiv:2007.07306*, 2020.

[8] Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. Temporally grounding natural sentence in video. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 162–171, 2018.

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[10] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[11] Zhenfang Chen, Lin Ma, Wenhan Luo, and Kwan-Yee K Wong. Weakly-supervised spatio-temporally grounding natural sentence in video. *arXiv preprint arXiv:1906.02549*, 2019.
[12] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.
[13] Samyak Datta, Karan Sikka, Anirban Roy, Karuna Ahuja, Devi Parikh, and Ajay Divakaran. Align2ground: Weakly supervised phrase grounding guided by image-caption alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2601–2610, 2019.
[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
[15] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482, 2015.
[16] Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould. Self-supervised video representation learning with odd-one-out networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3636–3645, 2017.
[17] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.
[18] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.
[19] Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. *arXiv preprint arXiv:2006.09920*, 2020.
[20] Tengda Han, Weidi Xie, and Andrew Zisserman. Memory-augmented dense predictive coding for video representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 312–329. Springer, 2020.
[21] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.
[22] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.
[23] Allan Jabri, Andrew Owens, and Alexei A Efros. Space-time correspondence as a contrastive random walk. *arXiv preprint arXiv:2006.14613*, 2020.
[24] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid. Action tubelet detector for spatio-temporal action localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4405–4413, 2017.
[25] Alicia Kirby. https://www.youtube.com/watch?v=r-ennr$_o h8a$.
[26] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.
[27] Zhenyang Li, Ran Tao, Efstratios Gavves, Cees GM Snoek, and Arnold WM Smeulders. Tracking by natural language specification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6495–6503, 2017.
[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.
[29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
[30] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889, 2020.
[31] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640, 2019.
[32] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
[33] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2019.
[34] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.
[35] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.
[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[37] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv*, 2018.
[38] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, 2015.
[39] Arka Sadhu, Kan Chen, and Ram Nevatia. Video object grounding using semantic roles in language description. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
[40] Ramprasaath R Selvaraju, Karan Desai, Justin Johnson, and Nikhil Naik. Casting your model: Learning to localize improves self-supervised representations. *arXiv preprint arXiv:2012.04630*, 2020.
[41] Dandan Shan, Jiaqi Geng, Michelle Shu, and David Fouhey. Understanding human hands in contact at internet scale. 2020.
[42] Jing Shi, Jia Xu, Boqing Gong, and Chenliang Xu. Not all frames are equal: Weakly-supervised video grounding with contextual similarity and visual clustering losses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10444–10452, 2019.
[43] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Contrastive bidirectional transformer for temporal representation learning. In *arXiv*, June 2019.
[44] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Learning video representations using contrastive bidirectional transformer. *arXiv preprint arXiv:1906.05743*, 2019.
[45] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.
[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.
[47] Jiangliu Wang, Jianbo Jiao, and Yun-Hui Liu. Self-supervised video representation learning by pace prediction. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, 2020.
[48] Josiah Wang and Lucia Specia. Phrase localization without paired training examples. In *Proceedings of the IEEE international conference on computer vision*, 2019.
[49] Donglai Wei, Joseph J Lim, Andrew Zisserman, and William T Freeman. Learning and using the arrow of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8052–8060, 2018.
[50] Woolworths. https://www.youtube.com/watch?v=viwpmylgps0.
[51] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *European Conference on Computer Vision*, 2018.
[52] Huijuan Xu, Abir Das, and Kate Saenko. R-c3d: Region convolutional 3d network for temporal activity detection. In *Proceedings of the IEEE international conference on computer vision*, pages 5783–5792, 2017.
[53] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*, 2019.
[54] Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10668–10677, 2020.
[55] Nanxuan Zhao, Zhirong Wu, Rynson WH Lau, and Stephen Lin. Distilling localization for self-supervised representation learning. *arXiv preprint arXiv:2004.06638*, 2020.
[56] Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J Corso, and Marcus Rohrbach. Grounded video description. In *CVPR*, 2019.
[57] Luowei Zhou, Nathan Louis, and Jason J Corso. Weakly-supervised video object grounding from text by loss weighting and object interaction. In *British Machine Vision Conference*, 2018.
[58] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.