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.

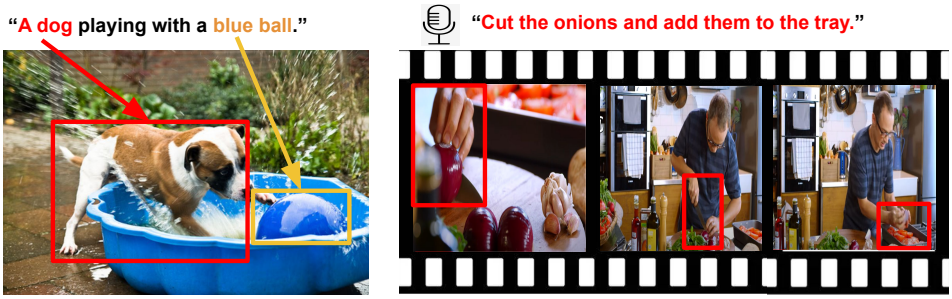
1 Introduction

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 [40]). 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 [57])

Contrastive learning [38] and computing contextual features via an attention model [52] 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 [35], 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 [22] or not using any paired text-image supervision at all [54]. 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 [36]. 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 [65] 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*. [22], and when combined with it, achieves state of the art on recall-at-one and pointing-hand accuracies.

2 Related Work

Self-supervised learning. There has been significant progress in natural language processing, resulting in effective and robust learned word representations [37, 17, 60, 39] 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. [10, 24, 11] have demonstrated that effective image representations can be learnt simply by contrasting between augmented views of the same image without labels. [62, 45] 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 [53, 23, 19, 26, 6, 55] 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) [51, 21, 20] as well as region-based convolutional neural network (CNNs) to detect objects or localize actions temporally [59] or spatio-temporally [27]. However, they are often trained on curated datasets and are, consequently, limited by a fixed predefined number of object categories in these datasets. [65] 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 [63]. These datasets provide bounding boxes for noun phrases [15, 63] or for egocentric videos [15]. 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 [29, 31, 40] 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 [35] or localizing relevant segments within untrimmed videos [5, 9, 12]. However, they do not identify the relevant spatial locations of the described interactions. Some existing works aim to spatially localize a natural language query [44, 61, 30] 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 [12, 4, 7]. To alleviate the costly annotations of curated datasets, [35] 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 [36]. In concurrent work, CLIP [41] 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 localization supervision, *e.g*., bounding boxes or segmentation masks. To this end, we propose a novel Contrastive Multilayer Multimodal Attention (CoMMA) module that allows for attention interactions between the features

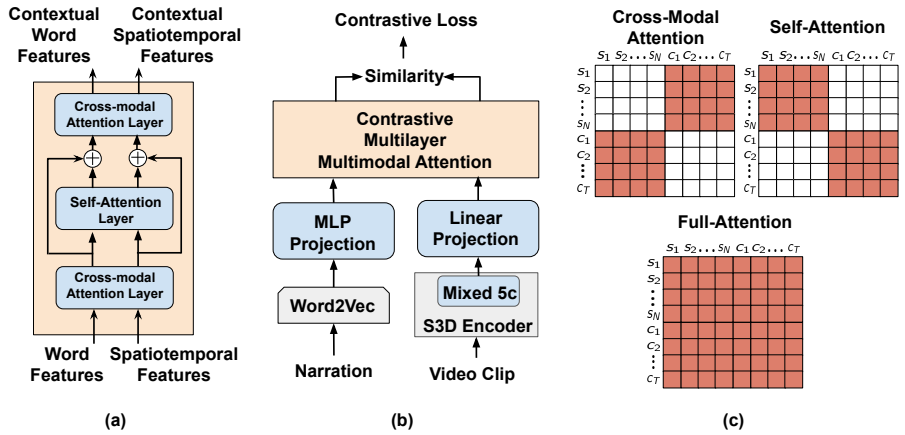


Figure 2: **Approach overview.** (a) Our proposed **Contrastive Multilayer Multimodal Attention (CoMMA)** module. (b) Our proposed architecture for computing joint attention over narrations and videos. (c) Binary masks for computing the different types of cross-modal attention. The shaded regions represent "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.

for spatiotemporal regions and words, as well as an effective contrasting of the two modalities' signals. We define two modalities to have attention interactions when their features attend to each other to compute their final contextualized representations. This setup is in contrast to when the two modalities' features only interact when the training loss is computed [35].

Our CoMMA module comprises alternating bidirectional cross-modal attention and self-attention layers (Figure 2(a)). The intuition underlying our proposed approach is that a hierarchical multimodal attention model with multiple stacked layers will encourage a fine-grained alignment between video spatiotemporal regions and narration words and ignore irrelevant regions/words. More specifically, each bidirectional cross-modal attention layer computes new representations for a target modality with the latent representations from the source modality to learn the relevance of each spatiotemporal 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. Through repeated attention interactions of the two modalities and aggregation of unimodal contextual information, our CoMMA module becomes more discriminative of irrelevant regions and words. Moreover, our module allows for applying a contrastive loss to the final representations for the two modalities since there is no cross-modal feature aggregation, *i.e.*, a modality may self-select features or select features from another modality by attending to them, but it may never select features from both modalities. Our approach goes beyond shallow attention architectures [3] and applying a contrastive loss to independent fixed-length features for each modality (*i.e.*, no attention interaction between the two modalities) [35]. Our multimodal attention module can be easily stacked on top of base feature encoders for our task (Figure 2(b) shows our full network). Next, we describe our CoMMA module in detail, how we employ a contrastive loss for training, and how we employ our network for the localization task.

3.1 Contrastive Multilayer Multimodal Attention

We formulate each attention layer in our CoMMA module as a key/query/value attention mechanism that is commonly used in attention models [52]. Our module accepts a concatenated sequence of narration word and video spatiotemporal region features and a set of binary attention masks. Let c_r be a spatiotemporal feature for region r and s_j be a word feature for the j-th word in a narration with N words. We obtain features c_r and s_j via base feature encoders 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_r and s_j, respectively, as column vectors.

In Figure 2(c), the shaded regions of the binary masks indicate which features interact during the attention computation for a layer. 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 setup differs

from the full-attention (FA) mask M^{FA} (illustrated in Figure 2(c) bottom) that computes attention over the *entire* multimodal sequence of words and regions and the self-attention (SA) mask M^{SA} (illustrated in Figure 2(c) top-right) that computes attention within each modality. We extend the mechanism of Vaswani *et al*. [52] by incorporating a mask that modulates the attention computation. Formally, let K, Q, V be the matrices of features for the keys, queries and values, respectively (the individual features are stacked as column vectors). Let the masked attention mechanism be represented as:

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

where M is any one of the binary masks shown in Figure 2(c), D denotes the scalar dimensionality of the query features, 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 $W^i_{K,l}$, $W^i_{Q,l}$ and $W^i_{V,l}$ 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 cross-modal attention layer by passing in the initial input features or contextualized feature outputs of the self-attention layer:

$$Y^{CA}_{l+1} = CrossAttn(Y^{SA}_l) = Attn(W^{CA}_{K,l+1}Y^{SA}_l, W^{CA}_{Q,l+1}Y^{SA}_l, W^{CA}_{V,l+1}Y^{SA}_l, M^{CA}) \tag{2}$$

where M^{CA} is the cross-modal attention mask. Note that if the input to Equation (2) is $Y^{SA}_l = [C_l, S_l]$, then the output will be $Y^{CA}_{l+1} = [S_{l+1}, C_{l+1}]$, *i.e.*, (contextualized) video features become (contextualized) language features and vice versa. 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. The similarity scores computed between the queries and the keys as the scaled dot product 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 to regions in the video clip 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.

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} ensures that attention is only computed between elements of the same modality. Formally, this computation can be represented as

$$Y^{SA}_l = SelfAttn(Y^{CA}_l) = Attn(W^{SA}_{K,l}Y^{CA}_l, W^{SA}_{Q,l}Y^{CA}_l, W^{SA}_{V,l}Y^{CA}_l, M^{SA}) + Y^{CA}_l \tag{3}$$

where M^{SA} is the self-attention mask. Its output is passed into the next cross-attention layer. We denote the final output of our module after L layers as C_L and S_L 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. Akbari et al. [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) [50]. 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(c)) and aggregates a summary knowledge into a sentinel vector. In contrast, CoMMA allows for repeated attention interactions without mixing of features from different modalities. We show in Section 4 that this mechanism is critical for localization.

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. $C_{L,r}$ denotes the column vector for the final contextualized feature for region r and $S_{L,j}$ denotes the column vector for the final contextualized representation for word j. The final video clip representation $\hat{C}$ and narration sentence representation $\hat{S}$ are computed by mean-pooling over the spatiotemporal regions $\hat{C} = \frac{1}{R}\sum_{r=1}^{R} C_{L,r}$ and words $\hat{S} = \frac{1}{N}\sum_{j=1}^{N} S_{L,j}$ respectively. Let $\left(\hat{C}^{(i)}, \hat{S}^{(i)}\right)$ be the i-th training example pair. We adapt the InfoNCE loss [38] by defining the sentence loss $\mathcal{L}_{sent}$ as the sum of log ratios over self-training pairs:

$$\mathcal{L}_{sent} = -\sum_{i=1}^{n} \log \left(\frac{\exp\left(\hat{C}^{(i)} \cdot \hat{S}^{(i)}\right)}{\exp\left(\hat{C}^{(i)} \cdot \hat{S}^{(i)}\right) + \sum\limits_{m \sim \mathcal{N}_{i,S}} \exp\left(\hat{C}^{(i)} \cdot \hat{S}^{(m)}\right) + \sum\limits_{m \sim \mathcal{N}_{i,C}} \exp\left(\hat{C}^{(m)} \cdot \hat{S}^{(i)}\right)} \right) \tag{4}$$

where the negative sets $\mathcal{N}_{i,C}$ and $\mathcal{N}_{i,S}$ comprise indices for non-corresponding video clip and narration pairs for the i-th training sample, and n denotes the total number of training samples. Note that our CoMMA module is applied to negative pairs as well. In contrast to the word-level objective (defined in Section 4) used in [22] which contrasts word features separately, we empirically find that contrasting aggregated features in the sentence loss $\mathcal{L}_{sent}$ better handles when the temporal alignment between narrations and videos is noisy.

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 $A = \prod_{l=0}^{L} W_l$ for attention weights W_l from the l-th layer. The resulting localization scores aggregate the total amount of attention by the entire set of spatiotemporal and word features assigned to a query feature.

4 Experiments

Word-level loss. We experiment with a word-level loss that is used in state-of-the-art phrase grounding models [22, 3], which aims to learn an alignment between each word in the narration and all spatial regions. Gupta *et al*. [22] has demonstrated that the word-level loss is effective at localizing noun phrases in images, which seeks to maximize the mutual information between each word and region features for corresponding video clip and narration pairs. We also consider an extension of the word-level loss where we incorporate the output contextualized representations from our CoMMA module. Let N denote the number of words in the narration, and n the total number of training samples. $S_{L,j}^{(i)}$ denotes the representation output of our CoMMA for the j-th word of the i-th training sample. Finally, for a given input word $S_{0,j}^{(i)}$, we compute its value representation with a multilayer perceptron (MLP): $\bar{S}_j^{(i)} = MLP(S_{0,j}^{(i)})$. Then, the word-level loss is formulated as the sum of log ratios:

$$\mathcal{L}_{word} = -\sum_{i=1}^{n}\sum_{j=1}^{N} \log \left(\frac{\exp\left(S_{L,j}^{(i)} \cdot \bar{S}_j^{(i)}\right)}{\exp\left(S_{L,j}^{(i)} \cdot \bar{S}_j^{(i)}\right) + \sum\limits_{m \sim \mathcal{N}_i} \exp\left(S_{L,j}^{(m)} \cdot \bar{S}_j^{(i)}\right)} \right) \tag{5}$$

where the negative set $\mathcal{N}_i$ of the i-th training sample comprises indices for the word features that are attended to by non-corresponding video clips.

4.1 Grounding interactions and objects in video

Self-training and evaluation datasets. We self-train our proposed model on instructional videos from the HowTo100M dataset [36]. 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. We use the publicly available base video and language feature encoders [35] that are trained on the entire HowTo100M dataset, which provides a good initialization for our learning task.

For evaluation, 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 [65]. We describe our dataset in full in the supplemental. Please see more details of the self-training and our collected dataset in the supplemental. We evaluate using our collected YouCook2-Interactions dataset for localizing narrated interactions.

In addition to localizing interactions given a sentence narration, we also consider localizing objects in video given a single object keyword. For localizing single objects, we evaluate on the YouCook2-BB dataset [64]. The YouCook2-BB dataset augments 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 criterion. Specifically, given 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 box. Otherwise, it is a "miss". The final localization accuracy is computed as the ratio of hits to the total number of hits and misses $\frac{\text{\# hits}}{\text{\# hits}+\text{\# misses}}$.

Implementation details. We use publicly available implementations for the separable 3D CNN-gated (S3D-G) encoder [58] and a shallow language encoder [35] built on top of Word2Vec [37] embeddings 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 optimizer [32] with one-epoch linear warmup. In our experiments, we also explore using two CoMMA modules but, due to the large memory requirements, the batch size has to be reduced significantly. 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.

During inference, 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.

Baselines. In our evaluations, we compare our approach against center prior, optical flow, and state-of-the-art image-ground baselines. In the center prior baseline, we return the center pixel for each frame. This 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. As most of the interactions involve motion in cooking videos, we compute an optical flow baseline. For this baseline, we generate an optical flow map for each frame using the OpenCV implementation of the Lucas-Kanade optical flow estimation method [8, 33]. Finally, we select the pixel with the highest magnitude as the location prediction.

Additionally, we compare against four cross-modal attention baselines. First, we compare against the model of Miech *et al*. [35] adapted to our task ("MIL-NCE"). We adapt their publicly available pretrained model by removing 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 are 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 computes full attention and leverages a sentinel vector ("Full attention + sentinel") to predict the similarity of the two modalities' features; an MLP is used to convert the output contextualized sentinel vector into a scalar MI-score which is then passed to an InfoNCE-like loss (inspired by CBT [49]). 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

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
MIL-NCE [35]	Sentence level	27.18
Contrastive phrase grounding [22]	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 report an ablation of the different attention module components in our model. We find that multiple stacked layers outperform shallower architectures. (Right) We ablate different types of self-attention over spatiotemporal features and find computing full self-attention over spatiotemporal features works best.

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

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

against a state-of-the-art phrase grounding model [22]. 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 task of interaction localization. Our proposed approach outperforms the center prior baseline by a significant margin of 20% on the task of localizing interactions given narration sentences. Surprisingly, computing the region with the maximum degree of motion through optical flow does not 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 alone is insufficient for our task. In addition, we observe that our proposed approach outperforms a state-of-the-art phrase grounding model [22], which is trained via optimizing the word-level loss (Equation 5), 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 module is trained with the word-level loss instead.

The difference in performance obtained by our model and the MIL-NCE baseline suggests that a model trained for retrieval may not be focusing on the relevant regions that are described by the sentence narration. 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, similar to Sun *et al*. [49], this baseline outputs a scalar value that is plugged in directly to the InfoNCE loss; it does not compute a cosine similarity between the video clip and natural language query contextualized features. We hypothesize that computing this MI-like score (instead of directly contrasting the two modalities' features) may somewhat mitigate this baseline from catastrophic failure where features from both modalities are aggregated during the attention 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 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 alignment between the different modalities. We report in Table 2 (right) results from an ablation of the different forms of self-attention over spatiotemporal regions. 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

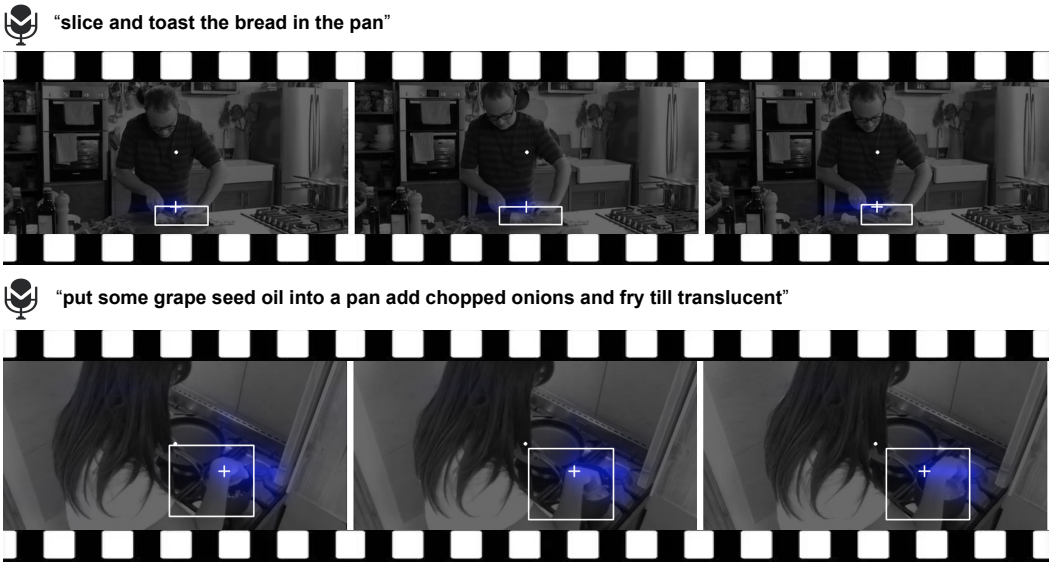


Figure 3: Examples of correct localization predictions by our model. Our approach localizes narrated interactions involving a person's hands and interacted objects. The boxes in these visualizations denote the ground-truth bounding boxes for the transcribed narrations. The white circle and cross indicate the locations of the center prior and the mode pixel of the computed attention score map, respectively. Finally, the blue regions indicate the regions that our CoMMA module determines to be most relevant to the transcribed narrations. (Video credit: Woolworths [57] and Alicia Kirby [28])

when self-attention is computed over all spatiotemporal regions. This finding suggests that reasoning about temporal context is important but it has to be done in conjunction with spatial context reasoning.

Object localization. Besides localizing full narrated sentences, we also evaluate our model's capability to localize single objects. Specifically, we compare our approach against the state-of-the-art "not all frames are equal" (NAFAE) model [47]. We report results in Table 3. Despite being trained with a sentence-level loss, we observe in Table 3 that our approach successfully localizes objects corresponding to single noun words. Notably, our proposed approach outperforms the NAFAE baseline. 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 score map 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 score map. 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 [47]	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 [22, 16, 18] by training on MSCOCO [31] and evaluating on Flickr30k [40].

We report results on the standard recall-at-K (R@K) and pointing hand recall accuracy criteria [22]. For training, we optimize the sum of the word-to-region matching loss and our sentence loss: $\mathcal{L}(\theta) = \mathcal{L}_{word} + \lambda \mathcal{L}_{sent}$. 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 [22] is extended with our CoMMA architecture as well as a description of the baselines are included in the supplemental. We note that the value projection layer for region features in [22] is omitted to incorporate CoMMA. Finally, we build on the same set of visual and language features from [22] for fair comparison.

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* [18]	VGG-cls (IN)	-	-	-	29.00
Akbari *et al* [3]	VGG-cls (IN)	-	-	-	61.66
Akbari *et al* [3]	PNAS Net (IN)	-	-	-	69.19
Align2Ground [16]	Faster-RCNN (VG)	-	-	-	71.00
Gupta *et al* [22]	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**

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 cross-attention layer. Adding a second cross-attention layer generally helps to improve the localization accuracy and leads to notable increases in R@1 and pointing hand recall accuracies, surpassing state-of-the-art [22] 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, we show in the supplemental that stacking multiple layers in our default CoMMA architecture when trained with the sentence-level loss $\mathcal{L}_{sent}$ 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 that this finding is a result of the captions consisting primarily of noun phrases and being strongly correlated with the images. 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 localization supervision. To address this task, we propose a novel Contrastive Multilayer Multimodal Attention module that facilitates repeated attention 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] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.
[9] 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.
[10] 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.
[11] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.
[12] 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.
[13] Collard Valley Cooks. https://www.youtube.com/watch?v=quunrafxqd0.
[14] Curiosity Culture. https://www.youtube.com/watch?v=jr9qw$_bjpte$.
[15] 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.
[16] 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.
[17] 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.
[18] 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.
[19] 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.
[20] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.
[21] 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.
[22] 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.
[23] 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.
[24] 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.
[25] 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.
[26] Allan Jabri, Andrew Owens, and Alexei A Efros. Space-time correspondence as a contrastive random walk. *arXiv preprint arXiv:2006.14613*, 2020.
[27] 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.
[28] Alicia Kirby. https://www.youtube.com/watch?v=r-ennr_oh8a.
[29] 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.
[30] 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.
[31] 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.
[32] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
[33] Bruce D Lucas, Takeo Kanade, et al. An iterative image registration technique with an application to stereo vision. Vancouver, British Columbia, 1981.
[34] Kharma Medic. https://www.youtube.com/watch?v=b$_zkdjlrghc$.
[35] 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.
[36] 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.
[37] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
[38] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2019.
[39] 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.
[40] 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.
[41] 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.
[42] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv*, 2018.
[43] 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.
[44] 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.
[45] 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.
[46] Dandan Shan, Jiaqi Geng, Michelle Shu, and David Fouhey. Understanding human hands in contact at internet scale. 2020.
[47] 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.
[48] Six Sisters' Stuff. https://www.youtube.com/watch?v=vic2l-iyrq0.
[49] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Contrastive bidirectional transformer for temporal representation learning. In *arXiv*, June 2019.
[50] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Learning video representations using contrastive bidirectional transformer. *arXiv preprint arXiv:1906.05743*, 2019.
[51] 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.
[52] 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.
[53] 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.
[54] Josiah Wang and Lucia Specia. Phrase localization without paired training examples. In *Proceedings of the IEEE international conference on computer vision*, 2019.
[55] 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.
[56] Smokin' Grillin' wit AB. https://www.youtube.com/watch?v=$_wvcrj15s-m$.
[57] Woolworths. https://www.youtube.com/watch?v=viwpmylgps0.
[58] 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.
[59] 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.
[60] 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.
[61] 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.
[62] Nanxuan Zhao, Zhirong Wu, Rynson WH Lau, and Stephen Lin. Distilling localization for self-supervised representation learning. *arXiv preprint arXiv:2004.06638*, 2020.
[63] Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J Corso, and Marcus Rohrbach. Grounded video description. In *CVPR*, 2019.
[64] 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.
[65] 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.

A Supplementary

In this supplementary material, we provide the following additions to the main submission:

A.1. weakly supervised phrase grounding experiments
  (a) implementation details
  (b) hybrid model
  (c) ablation results

A.2. grounding narrated instructions in videos experiments
  (a) self-training dataset
  (b) implementation details of baselines
  (c) mathematical formalization of cross-attention layer
  (d) additional details about our model architecture
  (e) Active Hands model implementation and results
  (f) limitations

A.3. details about our YouCook2-Interactions dataset
  (a) description of the annotation process
  (b) the guiding principles for frame relevance labeling
  (c) the guiding principles for bounding box annotation

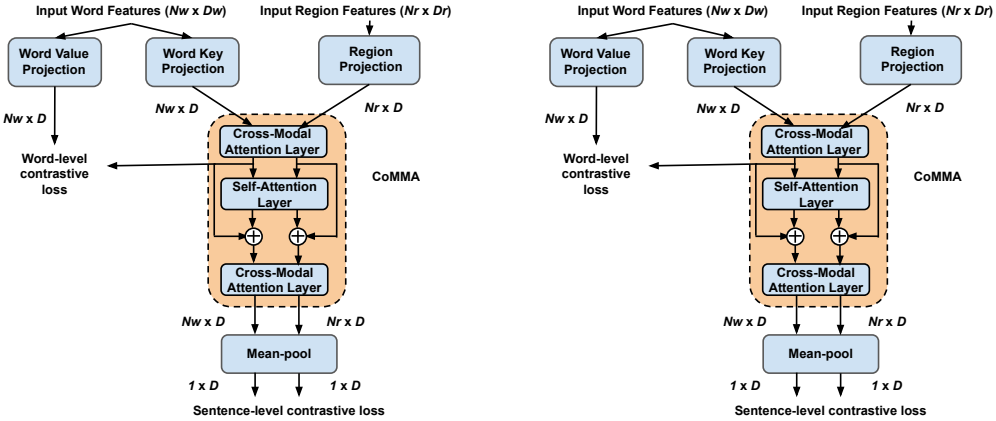
A.1 Weakly supervised phrase grounding experiments

Implementation details. Following Gupta *et al.* [22], we extract image region features with a Faster-RCNN [43] model pretrained on Visual Genome [29] and contextualized word embeddings from a pretrained BERT [17] model. The text value projection is implemented as a two-layer MLP with Batch Normalization [25]. We use ReLU as the activation function. We set the joint embedding space dimension to the same as Gupta *et al*. [22] (d=384) and use a batch size of 50 images. We adopt a constant learning rate of 1e-5 and the Adam optimizer. Our approach is built on the publicly available implementation of Gupta *et al*. [22]. Finally, we train our model for 8 hours on a single V100 GPU.

Hybrid model. We provide an illustration of our weakly supervised phrase grounding model in Figure 4b (this supplemental). D_w and D_r denote the dimensions of the input word and region features while N_w and N_r denote the number of words and regions, respectively. We use D to indicate the dimension of the joint embedding space. To incorporate our proposed CoMMA into the model of Gupta *et al*. [22] (Figure 4a (this supplemental)), we begin by removing a value projection MLP for the region features. This is due to the fact that our proposed CoMMA module does not project the region features to different subspaces for computing cross-attention. The model of Gupta *et al*. [22] is illustrated in Figure 4a (this supplemental). The outputs of the region and word key projection MLPs are passed as inputs into our cross-attention layer which computes contextualized region and word representations, respectively. We use the final contextualized word representations to compute the sentence-level loss. We found that using the contextualized word representation outputs from the first cross-attention layer to compute the word-level contrastive loss is critical for optimizing the model successfully. Finally, the sentence loss is weighted by a hyperparameter. We use the attention mask from the first cross-attention layer during evaluation.

Negative Noun Loss. In our experiments, we also adopt a negative noun loss as used in Gupta *et al*. [22]. Specifically, we create context-preserving negative captions for an image by substituting a noun in its original caption with negative nouns, that are sampled from a pretrained BERT [17] model. We compute the negative noun loss by contrasting the image region representations against these negative captions. For the rest of this supplementary, we use the term 'word-level contrastive loss' to include the negative noun loss as well.

Ablation Results. Table 5 (this supplemental) reports ablation results of the layers in our proposed CoMMA module when trained only with the sentence-level contrastive loss. In this case, adding cross-attention and self-attention layers generally helps to improve localization performance. As mentioned



(a) **Gupta *et al.*** [22]

(b) **Hybrid model for weakly supervised phrase grounding.**

Figure 4: N_w, N_r, D_w, D_r and D denote the number of words, regions as well as the dimensions of the input word and region features and the joint embedding space, respectively.

Table 5: **CoMMA component ablation on Flickr30k with only sentence-level loss.** We observe that adding cross-attention and self-attention layers generally help to improve the localization accuracy across the Recall@K criteria.

Cross-Modal Attention 1.	Cross-Modal Attention 2	Self-Attention	R@1	R@5	R@10	Pt Recall
✓	✗	✗	35.43	64.48	73.92	68.68
✓	✓	✗	39.03	66.91	75.15	**70.13**
✓	✓	✓	**39.31**	**68.43**	**77.35**	67.36

above, we remove the value projection for region features from the model of Gupta *et al*. [22] to incorporate the CoMMA module. This modification results in an initial drop in performance when a single cross-modal attention layer is used, as shown in the first row of Table 6 (this supplemental) and Table 4 (main paper). We also observe that the sentence-level contrastive loss is complementary to the word-level contrastive loss (on its own, it performs reasonably well). For example, in the case where only one cross-attention layer is used, adding the sentence-level contrastive loss leads to a 2.5% in the R@1 accuracy. However, using the complete CoMMA model hurts performance. We hypothesize that it is overfitting on the smaller dataset and we have not found the optimal optimization setting. Finally, Table 7 (this supplemental) reports the ablation results over the different weight values of the sentence-level contrastive loss. We achieve the best performance when we use a weight value of $\lambda = 0.005$ with two cross-attention layers.

A.2 Grounding narrated instructions in videos

Self-training dataset The HowTo100M [35] dataset contains approximately 1.2 million instructional videos that are sourced from YouTube, encompassing multiple domains such as cooking and hand-

Table 6: **Adding sentence loss to our hybrid model on Flickr30k.** We observe that the sentence loss is complementary to the word loss (on its own, it performs reasonably well). We also note that adding more attention layers also helps to improve localization performance.

Cross-Model Attention 1	Cross-Modal Attention 2	Self-Attention	Sentence Loss	R@1	R@5	R@10	Pt Recall
✓	✗	✗	No	49.99	**76.72**	**83.51**	74.97
✓	✗	✗	Yes	52.43	76.11	81.75	76.65
✓	✓	✗	Yes	**53.80**	76.69	82.28	**76.78**
✓	✓	✓	Yes	48.53	76.26	82.98	73.78

Table 7: **Sentence loss weight ablation with our hybrid model on Flickr30k.** The best performance is obtained with two cross-attention layers with the weight of the sentence loss set to 0.005. In the case where only one cross-attention layer is used, increasing the value of λ from 0 to 0.01 or 0.005 leads to a 2% in the R@1 accuracy.

Cross-Model Attention 1	Cross-Modal Attention 2	Self-Attention	Sentence Loss Weight λ	R@1	R@5	R@10	Pt Recall
✓	✗	✗	0.0	49.99	**76.72**	**83.51**	74.97
✓	✗	✗	0.01	52.59	75.42	80.82	75.80
✓	✗	✗	0.005	52.43	76.11	81.75	76.65
✓	✗	✗	0.001	45.72	71.05	77.84	71.99
✓	✓	✗	0.01	52.44	75.55	81.13	75.71
✓	✓	✗	0.005	**53.80**	76.69	82.28	**76.78**
✓	✓	✗	0.01	46.23	73.32	80.70	72.94
✓	✓	✓	0.01	47.57	75.59	82.63	72.92
✓	✓	✓	0.005	48.53	76.26	82.98	73.78
✓	✓	✓	0.01	48.81	76.33	82.99	75.32

crafting. These videos contain transcribed narrations that are either uploaded manually by users or are 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 [35], 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.

Implementation details. To avoid the time-consuming and computationally expensive process of pretraining the feature encoders used as inputs to our model, we leverage the publicly available pretrained weights of a S3DG model [35] trained with self-supervision. We use Word2vec embeddings [37] to represent semantic information contained by words in the narrations. Note that our proposed model is still fully self-supervised since the pretraining step for both the video and language encoders is also performed on the HowTo100M dataset without curated temporal or language annotations. The language encoder leverages pretrained Word2Vec [37] embeddings. In our experiments, we set the dimension of the joint embedding space D to be 512. We use identity and linear projections with the same input and output dimensions to represent W_K, W_Q and W_V in cross-attention and self-attention layers, respectively. To train our proposed model, we set a learning rate of 1e-4 and optimize the model using the AdamW [32] optimizer. At the beginning of training, we apply linear warmup for one epoch. We train our model on 8 V100 GPUs which takes about a day. During inference, we adopt the practice in prior work where the resolution of the input frame is set to 224 x 224. Each input video clip consists of 16 frames. The dimensions of the final feature map are T x H x W x D = 2 x 4 x 4 x 512, where T, H, W, D correspond to the temporal, height, width, and feature dimensions.

Mathematical formalization of the cross-attention layer We provide a more detailed formulation of the cross-attention layer in our CoMMA module here. In particular, we explain how we prevent the mixing of features from different modalities in this layer. Let us first consider an example where we have two features from different modalities x_1 and x_2. Let x be the concatenation of the two modalities' features $x = [x_1; x_2]$. We denote the operation in a linear layer as Ax for a matrix A. Let us write $A = [A_1\ A_2]$. Then, $Ax = A_1x_1 + A_2x_2$. In this case, the two modalities' features are fused since there is a summation operation between them.

Now, let us consider our setup where we have sets of "values" features V_1 and V_2 for two modalities. When full attention is used as in Figure 2c (bottom), the output of the softmax function in Equation (1) is a full weight matrix $A = [A_{11}\ A_{12}; A_{21}\ A_{22}]$. When multiplying with the value matrix $V = [V_1\ V_2]$ in Equation (1), the output is $[V_1A_{11} + V_2A_{21}\ V_1A_{21} + V_2A_{22}]$. Note that in this case, the feature sets for the two modalities (V_1 and V_2) are summed together.

When cross-attention is used as in Figure 2c (top-left), the output of the softmax function in Equation (1) is the matrix $A = [0\ A_{12}; A_{21}\ 0]$. When A is multiplied by the value matrix V, the output is $[V_2A_{21}\ V_1A_{12}]$. Consequently, the features for the two modalities are not summed together. The attention interaction between the two modalities only occurs when computing the softmax weights

A_{12} and A_{21}. We emphasize that these softmax weights are not features themselves, but instead are used for softly selecting features in the two modalities.

Implementation details of baseline architectures. Section 3.1 (main paper) describes existing state-of-the-art cross-modal attention modules. We provide additional implementation details of the full-attention baseline that uses a sentinel vector [49] and the shallow attention module [3]. The full-attention model differs from our proposed CoMMA module in two major ways. First, it computes attention weights over the entire multimodal sequence of words and clip features (full attention) instead of between modalities. Second, it uses a sentinel vector to aggregate information between visual and language features. The sentinel vector is a learnable parameter that has a similar purpose to the classification token that is commonly used in BERT language models [17]. Similar to its counterpart in the BERT model, the sentinel vector is prepended to the concatenated sequence of words and visual features and passed through a transformer to aggregate visual-semantic context in it.

The shallow attention module in Akbari *et al*.[3] computes a shallow cross-modal attention between words and convolutional features from different levels of an image encoder. The shallowness of its model stems from the fact that it only consists of a single layer of cross-modal attention between the language and visual features. In contrast, CoMMA comprises of alternating layers of cross-modal as well as self-attention layers. Note that this is different from the CBT model since it does not use a full attention module. As evidenced by our results, using a deeper cross-modal attention module improves the ability of our model to find the latent alignment between the language and visual modalities.

Finally, in Table 3 of the main paper, we note that we were unable to exactly reproduce the reported NAFAE results as the original base features were not available. Our best attempt with newer features was 4% lower than the reported bounding box detection accuracy. The final pointing accuracy obtained with the new features is 41.65%. Thus, we report the official higher bounding box detection accuracy of 46.95%. Our approach significantly outperforms this baseline.

Active Hands experiment implementations and results We compare our approach to a strongly supervised baseline that localizes *active hands* and interacted objects [46] in Table 8. We evaluate their pretrained model on YouCook2-Interactions under 3 settings:

1. taking the center pixel of the detected interacted object bounding box
2. taking the midpoint of the detected left and right hand bounding boxes
3. using a combination of the (1) and (2). In the case where no interacted objects are detected, we use (2) alone.

The high performance obtained by the pretrained Active Hands model is expected since hands are very often used in interactions, especially in the domain of instructional videos. However, they also suggest that simply detecting hands will likely not solve the proposed task. In particular, it will be challenging to resolve the ambiguity in scenes with multiple interactions involving different people that are occurring simultaneously. Additionally, this approach is not able to work well if there are erroneous detections of hands or objects or multiple pairs of hands visible. This is especially relevant in footage where the hands are small, making recognition based on local appearance difficult. Furthermore, detecting interacted objects using detected hands alone may not be feasible if the described interaction occurs between two objects without any human actors.

One of the main challenges with using the Active Hands model is that there are different heuristics that can be used to evaluate the localization accuracy, as reflected by the 3 settings described above. More importantly, this approach relies on large-scale bounding box annotations of hands during training and evaluation. In contrast, our proposed approach learns to localize such interactions from publicly available data without relying on such hand-annotated bounding boxes. Also, we would like to emphasize the generalizability of our proposed approach to the phrase grounding task in images that does not involve prominent human actors. We hypothesize that using an active hand detector will not generalize as well as our approach to this task given its reliance on finding noun phrases (instead of interactions) in images. Finally, we note that, in theory, it is possible to apply the CoMMA module on top of the hand and object bounding boxes to potentially improve the performance on the task in [46].

Limitations. One limitation is our method operates over coarse feature maps. Future work can aim to refine the localization predictions to compute precise masks for the interactions.

Table 8: **Interaction localization evaluation on YouCook2-Interactions.** We compare our approach against the pretrained strongly supervised Active Hands [46] model under 3 different settings.

Approach	Localization Accuracy (%)
Setting 1	70.40
Setting 2	61.47
Setting 3	64.47
CoMMA (Ours)	**55.80**

A.3 YouCook2-Interactions

Due to a lack of suitable datasets for localizing narrated interactions, we introduce an evaluation dataset, YouCook2-Interactions, that provides spatial bounding box annotations for interactions that are described by natural language sentences. Our dataset is built on the validation split of the YouCook2 dataset [65]. The original YouCook2 dataset contains around 2000 videos sourced from YouTube that cover approximately 89 recipes. It provides temporal annotations for segments and accompanying natural language descriptions. Each temporal segment is denoted by its start and end times and is described by a natural language description. Each description contains at least one interaction involving multiple objects (e.g., cut the chicken into cubes and add them to a pan).

A.3.1 Annotations

Our annotation process is split into three stages, which we will describe below. In the first stage (Person filtering), we filter frames using an off-the-shelf person detector. Next (Frame relevant labeling), we label each filtered frame as 'relevant' or 'irrelevant'. Finally (Bounding box annotations), we annotate each relevant frame with a spatial bounding box. We use qualified workers from the Amazon Mechanical Turk platform for both frame labeling and bounding box annotations.

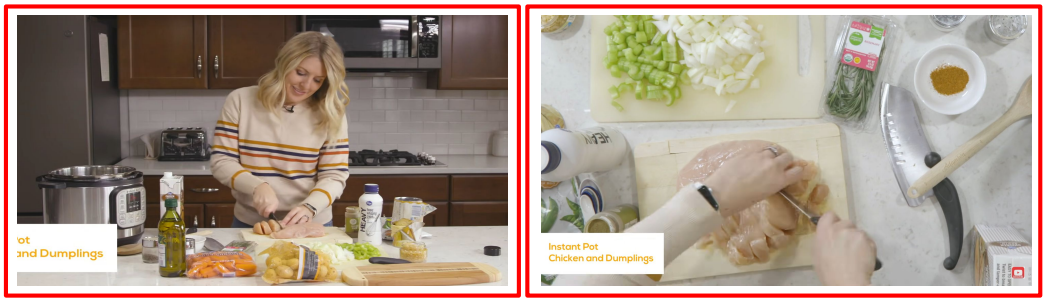
Person filtering. The primary focus of our dataset is on localizing narrated interactions. As such, frames which do not contain any instances of a person are removed to reduce the number of frames for labeling and annotation. To begin with, we extract all of the frames in the relevant segments from the validation split and run the YoloV3 object detector [42] on all of the frames to detect instances of humans. If a person is not detected in a given frame, it is removed from the labeling and annotation stages. Close-up frames that are visually relevant to narrated actions are not filtered out as long the object detector detects part of a person such as a hand.

Frame relevance labeling. One crucial characteristic of these narrations is that each specified interaction is only relevant to certain parts of the temporal segment. Unlike existing datasets which simply aim to track a referred object across frames, the specified interactions may not be visually relevant throughout the entire segment. Consequently, in the first stage, we ask workers to label if frames in a provided segment are relevant to any interactions specified within the description. We define people to be interacting with objects if they use their hands to perform an action involving the objects such as stirring or cutting. For each frame, three workers annotate whether the it is relevant and we select the majority label as the ground-truth label.

Bounding box annotations. After the set of relevant frames from each segment has been determined from the first stage of annotation, workers are tasked to draw a bounding box around the relevant region on a given frame that corresponds to the described interactions. We provide a general overview of the guiding principles here. See below for the entire set of principles. Since the primary objective of our approach is to localize interactions, the ideal bounding box annotation would enclose the region where the objects are being interacted with according to the natural language description. In cases where multiple objects are mentioned in the narration, in a given frame, only those that are being interacted with should be enclosed. Workers are also asked to enclose the entire hand and the object if the object is partially occluded in a frame.

The following sections include the guiding principles given to the annotators that are used in both the frame relevance labeling and bounding box annotation stages. For both stages, annotations are done on frames that are extracted at 1 frame per second (fps) and workers are asked to watch the entire clip to understand the context before they start annotating.

Sentence: add the chicken cubes into the instant pot



Explanation: both frames are irrelevant because cutting the chicken is not specified in the description.

Figure 5: Visual example of labeling frames as irrelevant because the sentence is not visually relevant. (Video credit: Six Sisters' Stuff [48])

A.3.2 Frame Relevance Labeling Guidelines

In the original temporal segment annotations provided in the YouCook2 dataset, we observed that the proportion of relevant frames within the specified temporal segments that corresponds to the described interactions is relatively small. To ensure that only relevant frames have bounding box annotations, the first step in our annotation process tasks workers to watch a given video clip and label if the frames are relevant. In this task, we define people to be interacting with objects if they use their hands to perform an action involving the objects such as stirring or cutting. Please note that if a person is just looking or pointing at the object, it is *NOT* considered as an 'interaction'.

A frame is labeled as '**irrelevant**' if it fulfills one of the following conditions:

1. it depicts a scene where the person would describe the objects or ingredients being used but the specified actions are not being carried out.
2. some sentences and their interactions may be ambiguous (*e.g.*, boiling water, bake the pizza in the oven). Frames where the person does not interact with the specified objects should be labeled as 'irrelevant'.
3. if the action is observed but the objects being interacted with are not visible at all. However, if parts of the objects are visible, then frame should be labeled as 'relevant' (Figure 9) (this supplemental).

As noted, there are ambiguous actions like 'cooking chicken in a pan', 'boiling water' or 'baking pizza in an oven'. In such cases, all frames with interacted objects as the action is being carried out should be labeled 'relevant'. Frames where the person does not interact with the object should be labeled as 'irrelevant'. We provide several visual examples with reasoning for their labels in Figures 5, 6 and 7 (this supplemental). The frames that are outlined in red and green are marked as 'irrelevant' and 'relevant', respectively.

A.3.3 Bounding Box Annotation Guidelines

Given the set of relevant frames that have been filtered by the first stage, Amazon Mechanical Turk workers are tasked to draw bounding boxes around the relevant regions given a natural language sentence. A visualization of the annotation interface can be found in Figure 8 (this supplemental). The general guiding principles are listed as follows:

1. Bounding boxes should only target the objects involved in the action as well as the entire hand / hands interacting with them. If only one hand is interacting with the object, the free hand should NOT be included.
2. The bounding box should only enclose the object(s) being interacted with, not all objects mentioned in the sentence.

Sentence: turn the corn cobs a few times as they are cooked

Explanation: The frame outlined in green is relevant since the person is in the process of flipping the corn cobs.

Figure 6: Visual example of labeling frames based on their visual relevance to the sentence. (Video credit: Collard Valley Cooks [13])

Sentence: pour the quinoa into the pan

Explanation: these frames are relevant since he is beginning the action of `pouring`.

Figure 7: Visual example of labeling frames based on their visual relevance to the sentence. (Video credit: Kharma Medic [34])

3. It should be as tight as possible.
4. In cases where the specified object is partially hidden, the bounding box should still enclose the entire hand and the object. For example in Figure 9 (this supplemental), some parts of the object are occluded by her hand. However, the bounding box still includes the entire hand.

Note that the HowTo100M dataset contains videos that overlap with those of YouCook2. To maintain a proper evaluation setting, overlapping videos are removed from the HowTo100M dataset. In our dataset, an interaction is defined as an action that involves the manipulation of at least one object.

A.4 Dataset statistics

We show dataset statistics in Table 9. In general, our dataset contains approximately 256 video segments that are split among 92 videos. Each video segment is denoted by its start and end times and corresponds to a single narration. The vocabulary encompasses 54 unique verbs and 73 object categories. Since the vocabulary is extracted from the YouCook2 dataset descriptions, they are mainly restricted to the culinary domain.

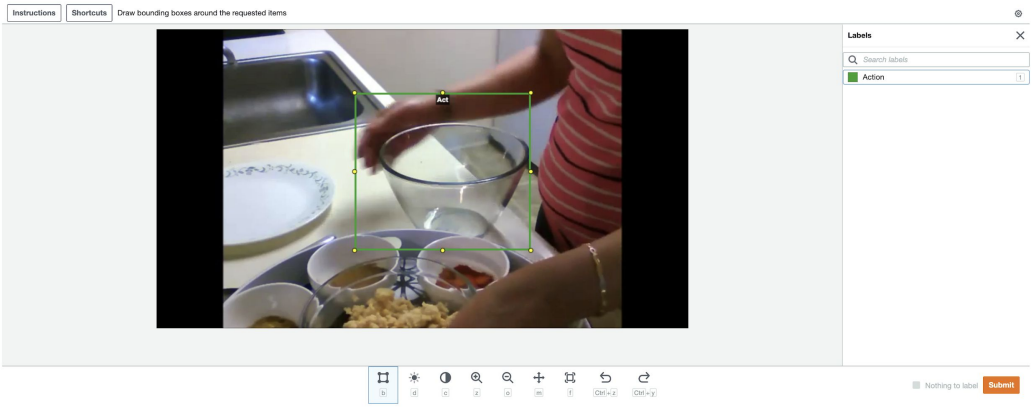


Figure 8: An example of the bounding box annotation interface used on Amazon Mechanical Turk. (Video credit: Curiosity Culture [14])

Figure 9: Visual example of how to annotate a bounding box when only part of the object is visible. (Video credit: Smokin' Grillin' wit AB [56])

# segments	256
# videos	92
Average segment length	24 sec.
# frames	6238
# verbs	54
# nouns	73

Table 9: **YouCook2-Interactions dataset statistics.** Notice that our collected evaluation dataset with bounding box annotations supports a range of described actions and objects.