What, when, and where? Self-Supervised Spatio-Temporal Grounding in Untrimmed Multi-Action Videos from Narrated Instructions

Brian Chen 1, Nina Shvetsova 2,3, Andrew Rouditchenko 4, Daniel Kondermann 5
Samuel Thomas 6,7, Shih-Fu Chang 1, Rogerio Feris 6,7, James Glass 4, Hilde Kuehne 2,3,7

1Columbia University, 2Goethe University Frankfurt, 3University of Bonn, 4MIT CSAIL
5Quality Match GmbH, 6IBM Research AI, 7MIT-IBM Watson AI Lab

{bc2754,sc250}@columbia.edu, shvetsov@uni-frankfurt.de, {roudi,glass}@mit.edu
dk@quality-match.com, {stomas,rsferis}@us.ibm.com, kuehne@cs.uni-bonn.de

Abstract

Spatio-temporal grounding describes the task of localizing events in space and time, e.g., in video data, based on verbal descriptions only. Models for this task are usually trained with human-annotated sentences and bounding box supervision. This work addresses this task from a multimodal supervision perspective, proposing a framework for spatio-temporal action grounding trained on loose video and subtitle supervision only, without human annotation. To this end, we combine local representation learning, which focuses on leveraging fine-grained spatial information, with a global representation encoding that captures higher-level representations and incorporates both in a joint approach. To evaluate this challenging task in a real-life setting, a new benchmark dataset is proposed, providing dense spatio-temporal grounding annotations in long, untrimmed, multi-action instructional videos for over 5K events. We evaluate the proposed approach and other methods on the proposed and standard downstream tasks, showing that our method improves over current baselines in various settings, including spatial, temporal, and untrimmed multi-action spatio-temporal grounding.

1. Introduction

Spatio-temporal grounding (STG) describes the challenging task of locating events in space and time within video data based on text referential expressions. Methods in this field usually rely on a combination of spatio-temporal bounding box annotation, together with a human-generated caption, describing the visual content of the bounding box [23, 54], which limits their generalizability beyond the given training scenario. Compared to that, as a second line of work, multimodal self-supervised learning tries to leverage “free” data sources, such as video and automatic speech recognition (ASR) captions from large-scale instructional videos to learn representations without human annotation [3, 4, 8, 35, 36]. The resulting models achieve state-of-the-art performance on zero-shot tasks such as cross-modal video retrieval or classification and also for zero-shot temporal action segmentation and detection based on free text queries [8, 28, 42, 47, 66], but usually lack spatial localization abilities. A third line of work focuses on label-free spatial grounding, e.g. by training on image-caption...
The goal is to correctly localize a referential expression in an image or each video frame, e.g., via a bounding box or a heatmap. However, those methods are not optimized to detect whether an event is present in a video. The assumption is thus that the evaluated expression is visible in the image or in all video frames.

The following work aims to bring together those ideas to address the task of spatio-temporal action grounding from multimodal supervision in untrimmed videos. We propose a grounding approach that uses video-text pairs based on ASR transcripts in instructional videos and learns the spatial representation of free-text events as well as their temporal extent, as shown in Figure 1. To this end, we leverage two different representations of the visual data: a global feature representation based on full-frame information to define the temporal extent of an event and a local representation based on frame-wise grid features for spatial localization. The motivation for this dualism is that while the local representation captures the spatial correlations between vision and text input, this can be too fine-grained to learn a holistic representation of the frame, while the global representation can be assumed to capture a more compact, aggregated view compared to local data and thus to provide a more reliable cue for the task of temporal localization. However, compared to the hand-annotated video-caption setup of most spatio-temporal grounding methods, the ASR text can be noisy as not all comments refer to visible events. Further, as there is only a loose temporal correlation, the described activities might not be precisely aligned, can be scattered over multiple frames, or not be present at all [18, 36]. Therefore, we propose to specifically select frames to capture only those useful for training. To this end, we look for frames that match the vocabulary of the respective text, leveraging a selection strategy by Sinkhorn optimal transport [11]. This allows us to train a model that can localize actions in space and time within videos without labeling supervision.

To evaluate spatio-temporal grounding in untrimmed videos, a new benchmark, GroundingYouTube, is proposed. It is based on the existing MiningYouTube dataset [28] and extended with spatio-temporal localization information. This setup differs from other benchmarks such as [10, 46, 63] in two ways: first, by using multiple center point annotations, it focuses on the grounding of referential actions itself instead of interacting humans or objects which are usually labeled; second, the dense annotations of multiple actions in the video allow us to benchmark action grounding in long, realistic untrimmed videos compared to existing, often pre-clipped benchmarks [10, 59]. The benchmark provides queries for 512 different event types and over 5K spatio-temporal annotations, as shown in Figure 1. A comparison of current datasets is shown in Table 1.

To evaluate the proposed approach as well as the new benchmark, the system is trained on the HowTo100M dataset [35] and compared to state-of-the-art methods based on full, weak, and self-supervision for spatial and temporal, as well as combined spatio-temporal grounding tasks. It shows that existing methods usually do well in one of the two aspects, spatial or temporal grounding. In contrast, the proposed method can combine spatial and temporal aspects.

We summarize the contributions of this work as follows:

1. We propose a framework for spatio-temporal grounding in untrimmed videos based on weakly aligned multimodal supervision without human annotation, employing a combination of global and local representation learning to learn the spatio-temporal extent of actions. (2) To facilitate this task, we propose a frame selection strategy based on Sinkhorn-Knopp Optimal transport that improves the quality of the acquired learning samples, leading to more effective supervision. (3) We provide a new benchmark and annotations to evaluate this challenging problem on real-world multi-action instructional video data.

### 2. Related Work

#### Supervised Spatio-temporal Grounding

Spatio-temporal Grounding refers to the problem of localizing a sequence of bounding boxes (a spatio-temporal tube) for a target object described by an input text. This problem has been addressed by various approaches TubeDETR [54], STCAT [23], STVGBert [45], STVG Bert [45].
STGVT [48], STGRN [59], Visil [27]. These methods rely on proposal networks such as Faster R-CNN [39] or MDETR [25] to predict bounding box coordinates for learning text-to-region interaction. All those approaches rely on supervised training with the human-annotated sentence and bounding box supervision, provided, e.g., by datasets such as VidSTG [59], HC-STVG [10]. While those datasets provide a temporal aspect, temporal detection is usually limited to identifying the start and end frame of a single action in a video. Compared to that, an untrimmed setting usually comprises multiple actions in a longer video that can be separated by longer background sequences. This conceptually differs from previous works [10] that typically use short videos of around 5-10 seconds. Other datasets such as ActivityNet entities [65] provide only bounding boxes for noun phrases in the captions, namely the objects, which is related to object detection task and does not capture any spatial or temporal extent of actions.

Multimodal Self-supervised Learning. The field of multimodal self-supervised learning aims to learn data representations by leveraging large amounts of unlabeled data with multiple modalities. Early works [13, 52] started by projecting images and text into a joint visual-language embedding space, where embeddings of semantically similar pairs are close. Those ideas have now grown into systems such as MIL-NCE [36] using the HowTo100M dataset [35] to train a video-language embedding space from 1.2 million instructional videos paired with text descriptions from ASR. Follow-up works, including [3, 4, 8, 40, 42] show that using videos without annotation enables an effective multimodal embedding space via contrastive learning.

Based on those advantages, approaches started to address the problem of Spatial Video Grounding from multimodal self-supervised aiming to identify spatial locations in a trimmed video based on text descriptions without the need for bounding box annotation during training. One of the early works studied this task in the context of weakly supervised learning where we learn grounding with human-annotated captions of the video [63]. In this context, works [41, 46] have focused on object grounding benchmarks such as YouCook2-BoundingBox [64], which provides bounding box annotations for visible objects in cooking videos. Other works such as GLIP [31], RegionCLIP [62], and others [56, 58] combine the principles of large-scale vision-language training with bounding box fine-tuning on object detection datasets [16, 34]. Recently, the YouCook-Interactions dataset [46] and CoMMA [46] have been proposed for the spatial grounding of objects and actions with multimodal self-supervision from HowTo100M videos. These works assume that the video is temporally clipped with respect to the grounding phrase.

Compared to that, Temporal Video Grounding aims to determine the set of consecutive frames corresponding to a text query in an untrimmed video [9, 22, 43], thus predicting temporal boundaries of action instances. Recent work such as MIL-NCE [36], MCN [8], and VideoCLIP [33] utilize large-scale pretraining for grounding actions temporally via text-to-frame similarity on video datasets such as MiningYouTube [28] or CrossTask [66] without proposals. However, these methods lack spatial localization ability [57, 61].

3. A Global-Local Framework for Spatio-Temporal Grounding

3.1. General setup

The goal of the proposed method is to temporally and spatially localize actions based on free-text queries in untrimmed videos. To this end, two representations are learned, a local and a global one. We start with narrated video clips, each associated with a corresponding visual representation and text narration. Namely, for each clip \( \mathcal{X} = \{V, S\} \), let \( V \) stand for the video clip and \( S \) for the text narration sentence generated by the automatic speech recognition (ASR) system. Each clip \( V \) consists of \( U \times N \) spatio-temporal tokens \( \{v_{u,n}\} \), where \( u \in \{1, ..., U\} \) represents the number of frames in the video and \( n \in \{1, ..., N\} \) represents the number of spatial grid region tokens or features in a frame. The text sentence \( S \) consists of \( K \) words \( \{s_1, ..., s_K\} \). We represent localized features by the tokens from each modality, and the global features \( V, S \) are acquired either by mean-pooling over the local features (S3D) or by using the [CLS] token from the transformer (CLIP) as in Radford et al. [38]. We learn transformations \( f : V \to \mathbb{R}^d \) to a \( d \)-dimensional representation \( f(V) \in \mathbb{R}^d \) from the global representation \( V \), and \( g : S \to \mathbb{R}^d \), to produce similar \( d \)-dimensional text global embeddings: \( g(S) \in \mathbb{R}^d \). Similar to \( \{f, g\} \), we note \( \{f, g\} \) to be the transform for localized features, where local features \( \{v, s\} \) are also projected as \( d \)-dimensional representations.

3.2. Representation guided frame sampling

Learning from multimodal self-supervision is challenging since the narration is likely to be noisy, thus containing more information than the actual task descriptions due to poor temporal alignment or cut scenes [18], which is the key differences between weakly supervised vision-captions grounding and multimodal self-supervised grounding. This work pursues a frame selection strategy to improve object grounding and temporal alignment during training. We start from a longer sequence \( U \), where \( U > T \), which includes the video frames before and after the ASR boundaries that could contain actions or objects in the sentence. Our goal is to find \( T \) frames out of the \( U \) frames that are most relevant to the actions and objects in the sentence \( S \). We formalize it as an optimal transport problem utilizing the Sinkhorn-Knopp algorithm [11].
Optimal transport for text-to-frame assignment. To acquire the optimal assignment from word features to video frames, an assignment matrix \( Q \) is computed from each video and ASR pair as shown in Figure 2(a). This cross-model optimal transport mechanism is applied to assignment \( Q \) from the projected cross-model similarity \( P \) between word tokens and each video frame, where \( P = g(S) \otimes f(V) \in \mathbb{R}^{K \times U} \). To compute the assignment matrix, the text and video projection layers from the global representation in Figure 2(c) are used to project multimodal features into a common space for feature similarity calculation. We investigate various granularities of the features where we compute the similarity between the text features at the word (local) / sentence (global) level and the visual feature at frames (global) / spatiotemporal tokens (local) level to acquire \( P \), as shown in Table 5. To ensure that the word-to-frame assignment contains more diversity instead of just saturated assignments to a single video frame, we add a constraint by Sinkhorn that requires label assignments to be equally distributed across various video frames representing diverse object/action concepts. Details of the Sinkhorn optimal transport are included in the appendix 7.1.

3.3. Local representations for spatial localization

To capture multimodal interaction with finer granularity, we need to learn the projection between tokenized features as shown in Figure 2(d). We extract spatio-temporal region features \( v_{tn} \) from the video. Also, we extract word features \( s_k \) which represents the feature from word \( k \). All tokenized features are projected through a linear layer. To compute attention between the tokenized features, we stacked two cross-modal attention layers with a self-attention layer in the middle, as illustrated in Figure 2(d). Cross-modal attention is computed similar to the standard attention mechanism [29].

Given a spatio-temporal token \( v_{tn} \) from a video, we compute the attention score to all of the words \( s_k \), where \( k \in \{1, \ldots, K\} \) in the ASR sentence \( S \) by \( \alpha_{tnk} = \frac{\exp(\epsilon_{tnk})}{\sum_{k=1}^{K} \exp(\epsilon_{tnk})} \) in the same video clip, where \( \epsilon_{tnk} = \cos(\text{word}(v_{tn}), \text{word}(s_k)) \). We then acquire a context video token feature \( \tilde{v}_{tn} = \sum_{k=1}^{K} \alpha_{tnk} s_k \), which encoded text contextual information. Note that the contextual vector is represented by aggregating the representations from the other modality. Follow the standard self-attention computation [49] \( K, Q, V \) represent the features for the keys, queries, and values as: \( \text{Attn}(K, Q, V) = \text{softmax} \left( \frac{(Q^T K)}{\sqrt{d_k}} \right) V \) where \( d_k \) is the dimension of the key.

In our case, we feed each contextual feature \( \{\tilde{v}_{tn}, \tilde{s}_k\} \) right after the first cross-attention layer to the \( K, Q, V \) to acquire its self-attended representation. The localized attention model was trained using contrastive loss. To represent the video clip \( V \) and ASR sentence \( S \), we mean-pool over the spatio-temporal tokens in video \( \tilde{V} = \frac{1}{TN} \sum_{r=1}^{TN} \tilde{v}_r \), and words \( \tilde{S} = \frac{1}{K} \sum_{k=1}^{K} \tilde{s}_k \) respectively. Let \( (V^{(l)}, S^{(l)}) \) be the \( l \)-th training example pair. We adopt the Noise Contrastive Estimation (NCE) loss [17] and the localized attention losses \( \mathcal{L}_{\text{Local}} \):

\[
-\frac{B}{4} \sum_{l=1}^{B} \left[ \log \frac{e^{\tilde{v}_l \cdot \tilde{s}_l - \delta}}{e^{\tilde{v}_l \cdot \tilde{s}_l - \delta} + \sum_{k \neq l} e^{\tilde{v}_l \cdot \tilde{s}_k}} + \log \frac{e^{\tilde{v}_l \cdot \tilde{s}_l + \delta}}{e^{\tilde{v}_l \cdot \tilde{s}_l + \delta} + \sum_{k \neq l} e^{\tilde{v}_l \cdot \tilde{s}_k}} \right]
\]

(1)

where \( B \) is the batch. \( \tilde{V}^{imp}_k \) and \( \tilde{S}^{imp}_k \) represent imposter samples, and \( \delta \) is a margin hyperparameter.

3.4. Global representations for temporal

We learn the global representation of a video clip and a sentence by contrastive loss, as shown in Figure 2(c). We again use the NCE loss function [17]. The global contrastive loss \( \mathcal{L}_{\text{Global}} \) follows the formulation as Equation 1 while using the global representations \( V \) and \( S \), which are the [CLS] tokens or mean-pooled features from both modalities, instead of the local representations. Projecting the global features to the same space ensures that the features across different
modalities are comparable. Since global representations encode information from the entire video, it is essential in encoding temporal information for the later downstream tasks. The final model is optimized by the sum of both losses as $\mathcal{L}_{\text{Final}} = \mathcal{L}_{\text{Local}} + \mathcal{L}_{\text{Global}}$.

Figure 3. Spatio-temporal inference. Both representations are used for spatio-temporal grounding: Starting by predicting the action boundary, spatial grounding is performed on the selected frames using the predicted label to find corresponding regions.

3.5. Inference for spatio-temporal grounding.

To perform spatio-temporal grounding on untrimmed videos, we start from temporal action detection as shown in Figure 3. Given a pool of possible action descriptions on the left and an untrimmed video, we perform feature similarity matching using the global representation ([CLS] token or mean-pooled feature) per frame with a threshold $\tau$ to filter backgrounds. We pick the action class with the largest similarity score per frame. Later, we use the predicted action class and feed it into the local representation branch to compute spatial grounding. We follow [1] to compute feature similarity between visual tokens and text tokens through the cross-attention and self-attention. In the end, we acquire an attention heatmap for later downstream evaluation. More information on inferencing are in appendix 8.3.

4. GroundingYoutube Benchmark

Current downstream datasets either provide spatial [46], or temporal annotation [28], or spatio-temporal annotation [59] but only for short video clips with few frames before and after the action takes place. These datasets do not provide the opportunity to evaluate both aspects, spatial and temporal grounding, in an untrimmed long video manner. We, therefore, extend one of the current benchmarks, MiningYouTube [28], with 250 untrimmed videos with a duration of around 3-5 minutes and an average of 41.6 labeled action instances per video. Note that each video contains various classes. As the dataset already provides dense temporal annotations, we annotate the respective temporal action segments in the dataset with spatial information.

Annotating the spatio-temporal extent of actions can be challenging as there is no clear visible outline as, e.g., in object annotation, nor is there a unique signal to indicate the temporal begin and end points. Similarly, grounding systems do not produce pixel-exact bounding boxes, but rather indicate regions of interest. Detector-free spatial grounding models [5] address this fuzziness by relying on pointing game accuracy, thus only using the center point of the heat map for evaluation. Lending on this idea, annotators were asked to mark the presumed center point of the action. Compared to bounding boxes, center point annotation can be advantageous because annotators are not visually distracted by object outlines, so it is more likely that the most important region will be selected. We capture five annotations per frame, resulting in a density-based heatmap.

Starting from 5,091 clips showing one of the 512 action classes, we adopt the methodology used for temporal action localization developed in [15] and label one frame per second, resulting in 26,987 labeled frames. We annotated all frames with five repeats per image, resulting in five annotations per frame and 134,935 point labels in total. To evaluate using bounding boxes [24], we get the union of all annotated points with additional distance to construct the bounding box as shown in Figure 4. More information on the annotation process, bounding box derivation, and dataset analysis is provided in the appendix 10.

5. Experiments

5.1. Datasets

Training Data: HowTo100M dataset contains 1.2M instructional videos along with their corresponding automatically generated speech (ASR). The narrations may be inaccurate and do not always accurately depict the video scene.

Figure 4. Visualization of the point annotation and automatic bounding box generation from points. The red point represents the mean of the five annotation points. The points annotation captures diverse patterns in various action types.
 downstream tasks to evaluate spatio-temporal grounding abilities of various models (detailed description is included in the appendix 8.1). Details of the implementation and experimental settings can be found in the appendix 8.2. Inference setups for each baseline are described in Section 8.3.

### 5.3. Downstream Tasks

We considered the following downstream tasks to evaluate spatio-temporal grounding abilities of various models (detailed description is included in the appendix 8.4):

1. **Spatial pointing game**: Given a text description to localize the region in the trimmed video. It is evaluated using the pointing game accuracy. If the predicted point lies in the ground truth bounding box, the result counts as a “hit” and counts as “miss” otherwise. The final accuracy is calculated as a ratio between hits to the total number of predictions. We also report the mean average precision (mAP) following the settings from V-HICO [32].

2. **Temporal grounding** provides videos with the respective pool of action instructions were provided. The model needs to localize each action step in time (start-time/end-time) and space (location in the video) as described in Figure 3. We evaluate in two metrics: IoU+Pointing game combines the evaluation setting from the spatial grounding [2] and temporal grounding [28] metrics. For each video frame, the prediction is correct when the model predicts the correct action for the frame. Also, given the predicted action as a query, the maximum point of the heatmap aims to lie within the desired bounding box. We then compute the Intersection over Union (IoU) over all the predictions with the GT to acquire the final score. We also compute video mAP following previous evaluation [15], where we set IoU threshold between GT and predicted spatio-temporal tubes. A prediction is correct when it surpasses the IoU threshold. We then compute the mAP over all classes. We form a 3D prediction mask following Figure 3 and compute IoU between our 3D heatmap and 3D tube.

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>DataSet</th>
<th>Supervision</th>
<th>Modality</th>
<th>IoU+Point</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoMMA† [46]</td>
<td>S3D</td>
<td>HT250K</td>
<td>Self</td>
<td>VT</td>
<td>1.02</td>
<td>1.26</td>
</tr>
<tr>
<td>MIL-NCE [36]</td>
<td>S3D*</td>
<td>HT100M</td>
<td>Self</td>
<td>VT</td>
<td>4.67</td>
<td>15.11</td>
</tr>
<tr>
<td>Ours</td>
<td>S3D</td>
<td>HT100M</td>
<td>Self</td>
<td>VT</td>
<td>9.12</td>
<td>25.92</td>
</tr>
<tr>
<td>GLIP [31]</td>
<td>Swin-L*</td>
<td>Cap24M</td>
<td>Weak</td>
<td>IT</td>
<td>1.24</td>
<td>1.56</td>
</tr>
<tr>
<td>CoMMA‡ [46]</td>
<td>CLIP</td>
<td>HT100M</td>
<td>Self</td>
<td>VT</td>
<td>1.68</td>
<td>1.82</td>
</tr>
<tr>
<td>CLIP [38]</td>
<td>CLIP</td>
<td>HT100M</td>
<td>Self</td>
<td>IT</td>
<td>3.59</td>
<td>12.74</td>
</tr>
<tr>
<td>RegionCLIP [62]</td>
<td>ResNet-101*</td>
<td>CC3M</td>
<td>Weak</td>
<td>IT</td>
<td>5.65</td>
<td>16.78</td>
</tr>
<tr>
<td>Ours</td>
<td>CLIP</td>
<td>HT100M</td>
<td>Self</td>
<td>VT</td>
<td>10.09</td>
<td>26.63</td>
</tr>
<tr>
<td>Ours</td>
<td>CLIP*</td>
<td>HT100M</td>
<td>Self</td>
<td>VT</td>
<td>11.53</td>
<td>28.26</td>
</tr>
<tr>
<td>MIL-NCE(temp.)+RegionCLIP(spa.)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>VT</td>
<td>9.21</td>
<td>22.33</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Backbone</th>
<th>DataSet</th>
<th>Supervision</th>
<th>Modality</th>
<th>IoU+Point</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet-101*</td>
<td>CC3M</td>
<td>Weak</td>
<td>VT</td>
<td>1.68</td>
<td>1.82</td>
</tr>
<tr>
<td>S3D</td>
<td>HT250K</td>
<td>Self</td>
<td>VT</td>
<td>1.02</td>
<td>1.26</td>
</tr>
<tr>
<td>S3D*</td>
<td>HT100M</td>
<td>Self</td>
<td>VT</td>
<td>4.67</td>
<td>15.11</td>
</tr>
<tr>
<td>S3D</td>
<td>HT100M</td>
<td>Self</td>
<td>VT</td>
<td>9.12</td>
<td>25.92</td>
</tr>
<tr>
<td>CLIP</td>
<td>HT100M</td>
<td>Self</td>
<td>VT</td>
<td>1.68</td>
<td>1.82</td>
</tr>
<tr>
<td>CLIP*</td>
<td>HT100M</td>
<td>Self</td>
<td>VT</td>
<td>11.53</td>
<td>28.26</td>
</tr>
</tbody>
</table>

Table 2. Spatio-temporal grounding on GroundingYouTube full videos. The proposed model learns global representations encoding global information and spatial correspondences across modalities, achieving a better performance in spatio-temporal evaluation compared to models trained on only spatial or temporal grounding. (V: video, I: image, T: text.) † indicates finetuned backbone.
the proposed framework to other methods on the temporal) and local (spatial) representations. We show that the proposed method improves over all other methods in this setting, while our model outperforms other grounding methods. * indicates finetuned backbone.

Table 3. Video spatial grounding. We evaluate the accuracy of the pointing game and the mean average precision. We listed CNN-based methods on top and transformer-based methods in the middle. Models learning global representations (MIL-NCE, CLIP) don’t perform well on localization tasks, while our model outperforms other grounding methods. * indicates finetuned backbone.

Table 4. Temporal Grounding on MiningYoutube. * indicates finetuned backbone. Spatial-focused model CoMMA is not trained for temporal detection, which results in lower performance. In contrast, the proposed model combines global and local representation, resulting in better temporal localization than one alone.

5.4. Comparison with state-of-the-art methods

(i) Spatio-temporal grounding in untrimmed video: We first compare the proposed method with other approaches designed for spatial or temporal grounding in Table 2. It shows that models without specific loss designs for spatial grounding (MIL-NCE, CLIP) show good mAP scores but lower pointing game accuracy. Out of the two weakly supervised methods, GLIP [31] and Region-CLIP [62], trained with aligned image-text, Region-CLIP show significantly better performance in this setting, while both perform in a similar range in the spatial grounding scenario (see Table 3). We attribute this behavior to the fact that RegionCLIP distinguishes frames with relevant queries better from background than GLIP, leading to better temporal localization. We finally compare the strong baseline MIL-NCE+RegionCLIP, which combines two approaches specialized in temporal and spatial aspects, to our task. It shows that the proposed method improves over all other baselines underlining the need to incorporate global (temporal) and local (spatial) representations.

(ii) Spatial grounding: Second, we compare the performance of the proposed framework to other methods on the task of spatial grounding, including models with weak supervision, as well as models trained in a fully supervised setting in Table 3. In the instruction video domain (GYT and YC-Inter), the proposed approach achieves the best result among all weakly and self-supervised trained methods. In the general domain (V-HICO and Daly), the method also achieves competitive results, showing the generalizability of the model to other domains. Note that in the Daly dataset, the classes are verbs, which are not detectable by the object-focused model GLIP. Compared to their weakly trained counterparts, fully-supervised model (TubeDETR [54], STCAT [23]) achieve competitive performance in the general domain (V-HICO, Daly) and slightly lower performance in instruction domain (GYT, YC-Inter) due to the domain gap with respect to the training data.

(iii) Temporal grounding: We evaluate temporal grounding in Table 4. Here, it shows that global representations also profit from local representation learning. This hypothesis is further validated in the ablation studies in Table 6, where we ablate both losses for all three settings and show a consistent improvement in the joint loss formulation.

5.5. Ablation study

We perform ablation studies with respect to all three settings, spatio-temporal grounding, as well as spatial and temporal grounding alone, reporting performance for spatio-temporal grounding on GroundingYT using mAP with IoU@0.4, on temporal grounding using MiningYT to IoU, and on spatial grounding using YC-Inter. pointing game. Additional ablation are in appendix 9.3.

Frame selection strategy. We perform an ablation on the possible frame selection strategies for our method (Figure 2(b) and Section 3.2). In Table 5, None uses all frames within the ASR boundary (U = T) as our video training data. Global represents the [CLS] token in text and video. Local uses the words and spatio-temporal tokens. In the setting Sinkhorn was not applied, the top T frames with the highest similarity score were selected. When we set
Table 5. **Frame selection**: (a) Sinkhorn selection results in better supervision. (b) We investigate all possible combinations of global and local representations for frame selection similarity matching. We found keeping the local text representation is crucial, and a combination of local and global representation leads to the best spatio-temporal grounding result.

Table 6. **Loss ablations**: both losses contribute to the final loss, and the existence of global loss helps the localization task.

5.6. Qualitative results

We visualize our spatio-temporal result in Figure 5. For the GLIP model, we output the bounding box with the highest confidence score and visualize its center point. We found GLIP model focuses on the salient object while our model focuses more on human-object interaction.

6. Conclusion

We presented an approach for learning spatio-temporal grounding with self-supervision and a new dataset: GroundingYoutube annotations, where we densely annotate spatio-temporal points/boxes from untrimmed multi-action videos. Our method includes a frame selection mechanism that identifies frames with groundable objects to adapt the learning process for untrimmed videos. Furthermore, we jointly learn global representations, which capture temporal information, and local representations learning fine-grained multimodal interactions between video and text. We conducted extensive experiments and our approach shows state-of-the-art performance in spatio-temporal grounding, as well as temporal and spatial grounding alone.

Acknowledgments: Nina Shvetsova is supported by the German Federal Ministry of Education and Research (BMBF) project STCL-01IS22067. Andrew Rouditchenko is supported by an NDSEG Fellowship and by the MIT-IBM Watson AI Lab. We thank the MIT-Satori cluster team for their technical support.
References


Zongheng Tang, Yue Liu, Si Liu, Guanbin Li, Xiaojie Jin, Hongxu Jiang, Qian Yu, and Dong Xu. Human-centric


[66] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos. 2019. 1, 2, 3