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]

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

{bc2754,sc250}@columbia.edu, shvetsov@uni-frankfurt.de, {roudi,glass}@mit.edu
dk@quality-match.com, {sthomas,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.

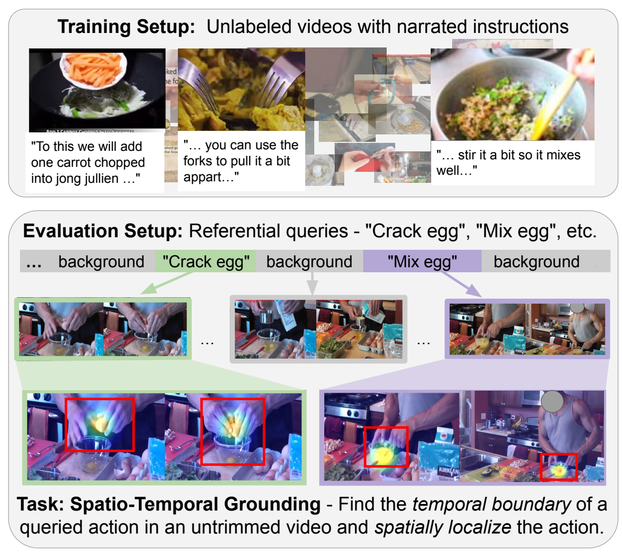

Figure 1. **Learning Spatio-temporal grounding in untrimmed videos.** In training, we learn from unlabeled videos without human annotation. In evaluation, we perform spatio-temporal grounding using an action description such as "crack egg" as a query. The model needs to localize both the action's temporal boundary and spatial region in the long untrimmed video. We visualize the heatmap from the annotation points as well as derived bounding boxes.

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

[2, 31, 50, 55, 62] or video-caption pairs [41, 46]. 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.

Dataset	Annotation	
	Spatial	Temporal
V-HiCo	object bb + human bb	-
AVA-Kinetics	object bb + human bb	-
ActivityNet entities	object bb + human bb	-
THUMOS14	-	action boundaries
ActivityNet	-	action boundaries
HACSSegment	-	action boundaries
YouCook II	-	multi-action boundaries
Cross-Task	-	multi-action boundaries
COIN	-	multi-action boundaries
EPIC KITCHENS-100	-	multi-action boundaries
Ego4D	-	multi-action boundaries
JHMDB51-21	human tubes	-
UCF101-24	human tubes	action boundaries
Daly	human tubes	action boundaries
Vid-STG	human tubes	action boundaries
HC-STVG	human tubes	action boundaries
AVA	human tubes	action boundaries
YouCook-Interactions	action bb	-
GroundingYoutube (ours)	action bb + center points	multi-action boundaries

Table 1. **Comparison of spatial, temporal, and spatio-temporal grounding datasets.** V-HiCo [32], AVA-Kinetics [30], THUMOS14 [20], ActivityNet [6], ActivityNet entity [65], HACSSegment [60], YouCook II [64], Cross-Task [66], COIN [47], EPIC KITCHENS-100 [12], Ego4D [14], JHMDB51-21 [21], UCF101-24 [44] Daly [51], Vid-STG [59], HC-STVG [48], AVA [15], YouCook-Interactions [46].

We summarize the contributions of this work as follows[1]: (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], STVGBert [45],

[1] We will make the code and the annotations publicly available.

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 $\mathcal{X} = \{\mathcal{V}, \mathcal{S}\}$, let $\mathcal{V}$ stand for the video clip and $\mathcal{S}$ for the text narration sentence generated by the automatic speech recognition (ASR) system. Each clip $\mathcal{V}$ 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 $\mathcal{S}$ 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 $f : V \rightarrow \mathbb{R}^d$ to a d-dimensional representation $f(V) \in \mathbb{R}^d$ from the global representation V, and $g : S \rightarrow \mathbb{R}^d$, 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-caption 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 $\mathcal{S}$. We formalize it as an optimal transport problem utilizing the Sinkhorn-Knopp algorithm [11].

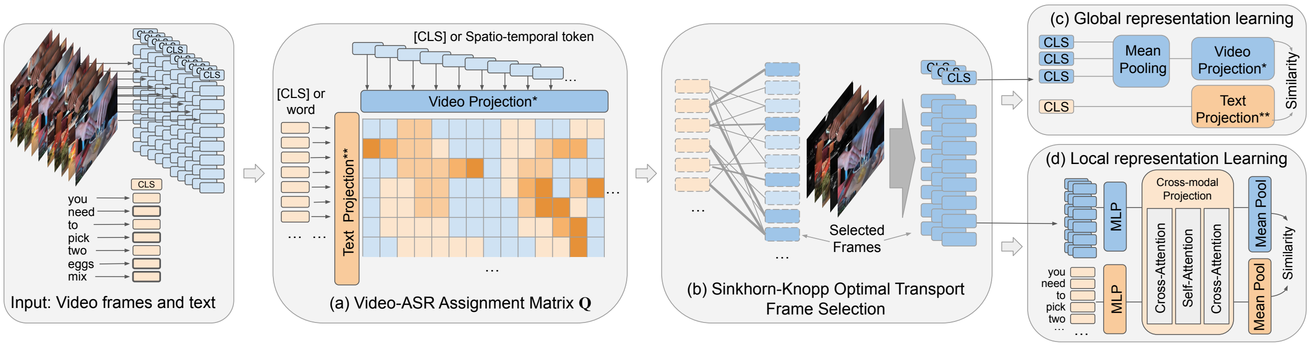

Figure 2. **Spatio-temporal grounding approach.** (a) We aim to select frames with groundable objects and actions. To this end, projected text features are matched with respective frame features. (b) Sinkhorn optimal transport is then leveraged to optimize the selected frames wrt. the text input. (c) Based on the selected frames, a global representation is learned to allow for temporal localization, as well as (d) a local representation to ground the action description in the spatial region.

Optimal transport for text-to-frame assignment. To acquire the optimal assignment from word features to video frames, an assignment matrix $\mathbf{Q}$ is computed from each video and ASR pair as shown in Figure 2(a). This cross-model optimal transport mechanism is applied to assignment $\mathbf{Q}$ from the projected cross-model similarity $\mathbf{P}$ between word tokens and each video frame, where $\mathbf{P} = g(\mathcal{S}) \bigotimes f(\mathcal{V})^\top \in \mathbb{R}^{K\times U}$. 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 $\mathbf{P}$, 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 $k \in \{1, ..., K\}$ in the ASR sentence $\mathcal{S}$ by $\alpha_{tnk} = \frac{\exp(e_{tnk})}{\sum_{k=1}^{K}\exp(e_{tnk})}$ in the same video clip, where $e_{tnk} = cosine(v_{tn}, s_k)$. We then acquire a contextual video token feature $\bar{v}_{tn} = \sum_{k=1}^{K}\alpha_{tnk}s_k$, 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: $Attn(K, Q, V) = \mathrm{softmax}\left(\frac{(Q^\top K)}{\sqrt{d_k}}\right)V$ where d_k is the dimension of the key. In our case, we feed each contextual feature $\{\bar{v}_{tn}, \bar{s}_k\}$ 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 $\mathcal{V}$ and ASR sentence $\mathcal{S}$, we mean-pool over the spatio-temporal tokens in video $\bar{V} = \frac{1}{TN}\sum_{r=1}^{TN}\bar{v}_r$, and words $\bar{S} = \frac{1}{K}\sum_{k=1}^{K}\bar{s}_k$ respectively. Let $\left(\bar{V}^{(l)}, \bar{S}^{(l)}\right)$ be the l-th training example pair. We adopt the Noise Contrastive Estimation (NCE) loss [17] and the localized attention losses $\mathcal{L}_{Local}$:

$$-\frac{1}{B}\sum_{l=1}^{B}\left[\left(\log\frac{e^{\bar{V}_l\cdot\bar{S}_l-\delta}}{e^{\bar{V}_l\cdot\bar{S}_l-\delta}+\sum\limits_{\substack{k=1\\k\neq l}}^{B} e^{\bar{V}_k^{imp}\cdot\bar{S}_l}}\right)+\left(\log\frac{e^{\bar{V}_l\cdot\bar{S}_l-\delta}}{e^{\bar{V}_l\cdot\bar{S}_l-\delta}+\sum\limits_{\substack{k=1\\k\neq l}}^{B} e^{\bar{V}_l\cdot\bar{S}_k^{imp}}}\right)\right] \tag{1}$$

where B is the batch. $\bar{V}_k^{imp}$ and $\bar{S}_k^{imp}$ represent imposter samples, and δ 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 $\mathcal{L}_{Global}$ 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 $\mathcal{L}_{Final} = \mathcal{L}_{Local} + \mathcal{L}_{Global}$.

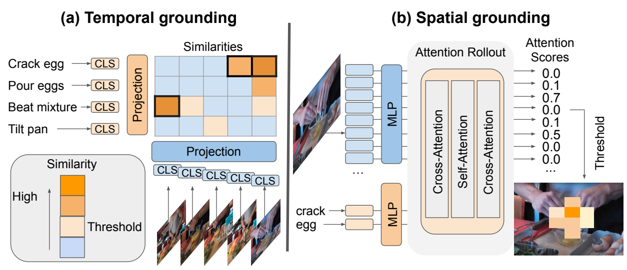

Figure 3. **Spatio-temporal inference.** Both representations are used for spatio-temporal grounding: Starting by predicting the action boundary, spatial grounding is performed on the selected frames using the predicted label to find corresponding regions.

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 τ 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

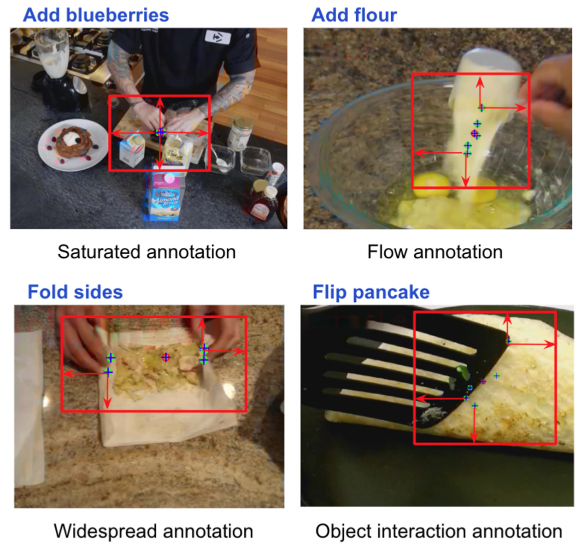

Figure 4. **Visualization of the point annotation and automatic bounding box generation from points.** The red point represents the mean of the five annotation points. The points annotation captures diverse patterns in various action types.

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.

Method	Backbone	DataSet	Supervision	Modality	GroundingYoutube IoU+Point	mAP 0.1	mAP 0.2	mAP 0.3	mAP 0.4	mAP 0.5	mAP 0.1:0.5
CoMMA† [46]	S3D	HT250K	Self	VT	1.02	2.18	1.72	1.11	0.93	0.37	1.26
MIL-NCE [36]	S3D*	HT100M	Self	VT	4.67	33.94	25.16	12.65	3.42	0.41	15.11
Ours	S3D	HT100M	Self	VT	**9.12**	**42.70**	**35.49**	**25.16**	**16.22**	**10.05**	**25.92**
GLIP [31]	Swin-L*	Cap24M	Weak	IT	1.24	2.83	2.10	1.52	0.96	0.37	1.56
CoMMA‡ [46]	CLIP	HT100M	Self	VT	1.68	3.51	2.32	1.88	0.99	0.40	1.82
CLIP [38]	CLIP	HT100M	Self	IT	3.59	29.54	22.15	9.16	2.48	0.39	12.74
RegionCLIP [62]	ResNet-101*	CC3M	Weak	IT	5.65	35.65	27.43	15.69	4.31	0.86	16.78
Ours	CLIP	HT100M	Self	VT	10.09	42.81	36.05	25.84	17.10	11.35	26.63
Ours	CLIP*	HT100M	Self	VT	**11.53**	**43.64**	**36.94**	**26.78**	**19.45**	**14.61**	**28.26**
MIL-NCE(temp.)+RegionCLIP(spa.)	-	-	-	VT	9.21	40.54	34.97	22.38	13.79	9.18	22.33

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.

Downstream Datasets: GroundingYoutube (GYT) is used to evaluate the task of multi-action spatio-temporal grounding as described in Section 4. **MiningYoutube (MYT)** [28] provides temporal annotation and is limited to the domain of cooking instruction videos. **YouCook-Interaction (YC-Inter)** [46] is an extension of the YouCook2 dataset [64] for cooking instruction providing bounding boxes for 6K selected frames. The bounding boxes usually comprise the hand and the tool mentioned in the respective sentence-wise annotation. To further benchmark on general video domains on the **V-HICO** dataset [32] with 6.5k videos with human-object interaction bounding boxes annotations, and **Daly** action dataset [51], featuring videos consisting of daily actions such as "brushing teeth".

5.2. Baseline methods

Temporal: MIL-NCE [36] utilizes S3D [53] and word2vec [37]. CLIP [38], an image-text model with transformer. **Spatial**: CoMMA [46], SSL model († for weights shared by the author[2] ‡ trained with CLIP); GLIP [31], RegionCLIP [62], SOTA weakly supervised grounding model. **Spatio-temporal**: We construct MIL-NCE+RegionCLIP following the inference pipeline in Figure 3. TubeDETR [54] and STCAT [23] are supervised. More descriptions of the baselines are given 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):

(i) **Spatio-temporal grounding in untrimmed video** is evaluated on the proposed Grounding Youtube dataset. The entire video and 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.

(ii) **Spatial grounding** is 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 $\frac{\#\text{ hits}}{\#\text{ hits}+\#\text{ misses}}$. We also report the mean average precision **(mAP)** following the settings from V-HICO [32].

(iii) **Temporal grounding** provides videos with the respective actions and their ordering, including the background. The goal is to find the correct frame-wise segmentation of the video. We follow the inference procedure in [28] to compute the alignment given the similarity input matrix. The task is evaluated by intersection over detection (IoD), defined as $\frac{G \cap D}{D}$ the ratio between the intersection of ground-truth action G and prediction D to prediction D, and the Jaccard index, which is an (IoU) given as $\frac{G \cap D}{G \cup D}$.

[2]We thank the authors for providing code and weights.

Method	Backbone	Data	Super.	Mod.	YC-Inter Acc	GroundingYT Acc	GroundingYT mAP	V-HICO Acc	V-HICO mAP	Daly Acc	Daly mAP
MIL-NCE [36]	S3D*	HT100M	Self	VT	23.67	27.45	8.21	12.65	11.23	13.84	24.23
CoMMA† [46]	S3D	HT250K	Self	VT	48.63	47.68	23.38	40.97	21.45	54.48	33.39
Ours	S3D	HT100M	Self	VT	**53.98**	**60.62**	**44.93**	**44.32**	**24.31**	**66.35**	**45.93**
CLIP [38]	CLIP	HT100M	Self	IT	14.10	12.50	3.49	29.23	12.51	18.02	27.28
CoMMA‡ [46]	CLIP	HT100M	Self	VT	52.65	47.56	36.42	55.20	34.54	61.06	44.37
RegionCLIP [62]	RN50x4*	CC3M	Weak	IT	51.56	52.84	23.42	57.92	37.82	67.12	48.62
GLIP [31]	Swin-L*	Cap24M	Weak	IT	52.84	53.62	24.73	**66.05**	41.17	-	-
Ours	CLIP	HT100M	Self	VT	57.10	55.49	43.12	60.71	39.28	70.08	50.56
Ours	CLIP*	HT100M	Self	VT	**58.35**	**56.98**	**45.32**	62.34	**41.56**	**71.35**	**52.78**
TubeDETR [54]	MDETR	Vid-STG	Full	VT	51.63	53.24	41.76	63.23	40.87	84.21	62.98
STCAT [23]	ResNet-101	Vid-STG	Full	VT	54.47	55.90	44.21	65.34	41.10	85.42	63.94

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.

Method	Backbone	Data	Super.	IoU	IoD
Mining: MLP [28]	TSM	MiningYT	Weak	9.80	19.20
CoMMA* [46]	S3D	HT250K	Self	2.05	5.63
RegionCLIP [62]	RN50x4*	CC3M	Weak	10.96	16.86
MIL-NCE [36]	S3D*	HT100M	Self	18.69	26.74
Ours	S3D	HT100M	Self	19.18	27.65
Ours	CLIP	HT100M	Self	19.88	28.50
Ours	CLIP*	HT100M	Self	**20.33**	**29.67**

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 [36], CLIP [38]) show good mAP scores but lower pointing game accuracy. Out of the two weakly supervised methods, GLIP [31] and RegionCLIP [62]), trained with aligned image-text, RegionCLIP 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 (TubeDETER [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 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

	GroundingYT Spatio-temporal	MiningYT Temporal	YC-Inter. Spatial	GroundingYT Spatio-temporal	MiningYT Temporal	YC-Inter. Spatial
	w/o Sinkhorn			with Sinkhorn		
None	17.43	18.34	57.42	-	-	-
Global T - Global V	18.24	19.38	56.31	18.96	19.89	57.34
Global T - Local V	18.01	19.31	57.56	18.53	19.35	57.67
Local T - Global V	18.05	19.85	57.34	19.32	**20.36**	58.13
Local T - Local V	18.31	19.48	57.86	19.43	20.16	**58.51**
Average over last two	18.36	19.68	57.77	**19.45**	20.33	58.35

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.

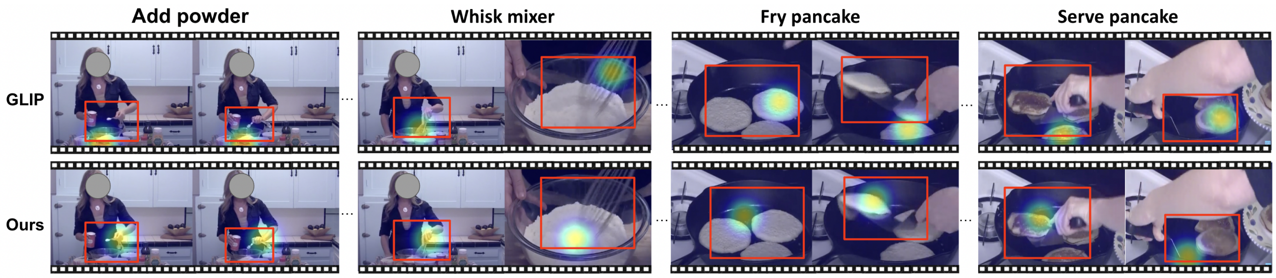

Figure 5. **Visualization on GroundingYoutube.** Red box: annotation. Heatmap: prediction. Baseline GLIP focuses on objects such as power, mixer, and pancake. Our method focuses on actions such as whisk and serve.

	GroundingYT Spatio-temp	MiningYT Temporal	YC-Inter. Spatial
only Local loss	7.29	5.23	55.29
only Global loss	9.28	19.12	36.23
w/ Both loss	19.45	20.33	58.35

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

spatio-temporal tokens as the selection target, we sum over the scores with respect to each frame to acquire the frame similarity score. It shows that selecting frames based on Sinkhorn selection leads to consistently better results as it enforces more variety of visual concepts but also captures frames with possible groundable objects. It further shows that word tokens are more suitable than the global text CLS token for frame selection. Finally, we see that depending on the task (spatial vs. temporal), a local resp. global representation is better, and a combination of both works best for spatio-temporal grounding. We provide runtime analysis of such frame selection strategy in the appendix 9.1.

Global and local loss. As mentioned in the spatio-temporal evaluation, both features contribute to the final grounding result. We test the model by ablating out each loss. Table 6 shows that each loss not only contributes to the spatio-temporal grounding on the GYT, but also that the whole is more than the sum of its parts (losses) since this task requires both spatial and temporal detection. The reduced impact of the global loss in the case of YC-Inter is that this is a pure spatial grounding dataset (no background frames) without temporal detection, and the local loss plays a more critical role. We observe the same patterns in the temporal grounding result for MYT, where spatial localization is not directly contributing to the final performance. We tried out the same ablation using in the S3D backbone in supplement.

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

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020. 5, 13

[2] 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. 2, 6, 13

[3] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34, 2021. 1, 3

[4] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. *Advances in Neural Information Processing Systems*, 33:25–37, 2020. 1, 3

[5] Assaf Arbelle, Sivan Doveh, Amit Alfassy, Joseph Shtok, Guy Lev, Eli Schwartz, Hilde Kuehne, Hila Barak Levi, Prasanna Sattigeri, Rameswar Panda, et al. Detector-free weakly supervised grounding by separation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1801–1812, 2021. 5

[6] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 2

[7] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020. 12

[8] Brian Chen, Andrew Rouditchenko, Kevin Duarte, Hilde Kuehne, Samuel Thomas, Angie Boggust, Rameswar Panda, Brian Kingsbury, Rogerio Feris, David Harwath, et al. Multimodal clustering networks for self-supervised learning from unlabeled videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8012–8021, 2021. 1, 3, 14

[9] Yi-Wen Chen, Yi-Hsuan Tsai, and Ming-Hsuan Yang. End-to-end multi-modal video temporal grounding. *Advances in Neural Information Processing Systems*, 34:28442–28453, 2021. 3

[10] Zhenfang Chen, Lin Ma, Wenhan Luo, and Kwan-Yee Kenneth Wong. Weakly-supervised spatio-temporally grounding natural sentence in video. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1884–1894, Florence, Italy, 2019. Association for Computational Linguistics. 2, 3

[11] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. 2013. 2, 3, 12

[12] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pages 1–23, 2022. 2

[13] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems*, 26, 2013. 3

[14] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 2

[15] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6047–6056, 2018. 2, 5, 6, 13, 14

[16] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 3

[17] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS*, 2010. 4

[18] Tengda Han, Weidi Xie, and Andrew Zisserman. Temporal alignment networks for long-term video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2906–2916, 2022. 2, 3

[19] David Harwath, Adria Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass. Jointly discovering visual objects and spoken words from raw sensory input. In *Proceedings of the European conference on computer vision (ECCV)*, pages 649–665, 2018. 14

[20] H. Idrees, A. R. Zamir, Y. Jiang, A. Gorban, I. Laptev, R. Sukthankar, and M. Shah. The thumos challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, 155:1–23, 2017. 2

[21] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. Towards understanding action recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 3192–3199, 2013. 2

[22] Yu-Gang Jiang, Jingen Liu, A Roshan Zamir, George Toderici, Ivan Laptev, Mubarak Shah, and Rahul Sukthankar. Thumos challenge: Action recognition with a large number of classes. http://crcv.ucf.edu/THUMOS14/, 2014. 3, 14

[23] Yang Jin, yongzhi li, Zehuan Yuan, and Yadong MU. Embracing consistency: A one-stage approach for spatio-temporal video grounding. In *Advances in Neural Information Processing Systems*, 2022. 1, 2, 6, 7, 14, 15

[24] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid. Action tubelet detector for spatio-temporal action localization. In *Proceedings of the IEEE

International Conference on Computer Vision, pages 4405–4413, 2017. 5

[25] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021. 3

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 2015. 13

[27] Giorgos Kordopatis-Zilos, Symeon Papadopoulos, Ioannis Patras, and Ioannis Kompatsiaris. Visil: Fine-grained spatio-temporal video similarity learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6351–6360, 2019. 3

[28] Hilde Kuehne, Ahsan Iqbal, Alexander Richard, and Juergen Gall. Mining youtube-a dataset for learning fine-grained action concepts from webly supervised video data. 2019. 1, 2, 3, 5, 6, 7, 13

[29] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, pages 201–216, 2018. 4

[30] Ang Li, Meghana Thotakuri, David A Ross, João Carreira, Alexander Vostrikov, and Andrew Zisserman. The ava-kinetics localized human actions video dataset. *arXiv preprint arXiv:2005.00214*, 2020. 2

[31] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 2, 3, 6, 7, 12, 15

[32] Shuang Li, Yilun Du, Antonio Torralba, Josef Sivic, and Bryan Russell. Weakly supervised human-object interaction detection in video via contrastive spatiotemporal regions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1845–1855, 2021. 2, 6, 13

[33] Yikang Li, Jenhao Hsiao, and Chiuman Ho. Videoclip: A cross-attention model for fast video-text retrieval task with image clip. In *Proceedings of the 2022 International Conference on Multimedia Retrieval*, pages 29–33, 2022. 3

[34] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 3

[35] 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. 1, 2, 3, 12

[36] 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. 1, 2, 3, 6, 7, 12, 15

[37] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *arXiv preprint arXiv:1301.3781*, 2013. 6, 12

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 3, 6, 7, 12, 15

[39] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 3

[40] Andrew Rouditchenko, Angie Boggust, David Harwath, Brian Chen, Dhiraj Joshi, Samuel Thomas, Kartik Audhkhasi, Hilde Kuehne, Rameswar Panda, Rogerio Feris, et al. Avlnet: Learning audio-visual language representations from instructional videos. *arXiv preprint arXiv:2006.09199*, 2020. 3

[41] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 10444–10452, 2019. 2, 3

[42] Nina Shvetsova, Brian Chen, Andrew Rouditchenko, Samuel Thomas, Brian Kingsbury, Rogerio Feris, David Harwath, James Glass, and Hilde Kuehne. Everything at once–multi-modal fusion transformer for video retrieval. In *CVPR*, 2022. 1, 3, 14

[43] Mattia Soldan, Alejandro Pardo, Juan León Alcázar, Fabian Caba Heilbron, Chen Zhao, Silvio Giancola, and Bernard Ghanem. Mad: A scalable dataset for language grounding in videos from movie audio descriptions. *arXiv preprint arXiv:2112.00431*, 2021. 3

[44] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2

[45] Rui Su, Qian Yu, and Dong Xu. Stvgbert: A visual-linguistic transformer based framework for spatio-temporal video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1533–1542, 2021. 2

[46] Reuben Tan, Bryan Plummer, Kate Saenko, Hailin Jin, and Bryan Russell. Look at what i'm doing: Self-supervised spatial grounding of narrations in instructional videos. *Advances in Neural Information Processing Systems*, 34, 2021. 2, 3, 5, 6, 7, 12, 15

[47] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1207–1216, 2019. 1, 2

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

spatio-temporal video grounding with visual transformers. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021. 2, 3

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[50] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11686–11695, 2022. 2

[51] Philippe Weinzaepfel, Xavier Martin, and Cordelia Schmid. Human action localization with sparse spatial supervision. *arXiv preprint arXiv:1605.05197*, 2016. 2, 6

[52] Jason Weston, Samy Bengio, and Nicolas Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011. 3

[53] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 305–321, 2018. 6, 12

[54] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Tubedetr: Spatio-temporal video grounding with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16442–16453, 2022. 1, 2, 6, 7, 14, 15

[55] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. Unitab: Unifying text and box outputs for grounded vision-language modeling. In *European Conference on Computer Vision*, pages 521–539. Springer, 2022. 2

[56] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *Advances in Neural Information Processing Systems*, 35:9125–9138, 2022. 3

[57] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. Dense regression network for video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10287–10296, 2020. 3

[58] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. *Advances in Neural Information Processing Systems*, 35:36067–36080, 2022. 3

[59] 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. 2, 3, 5

[60] Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. Hacs: Human action clips and segments dataset for recognition and temporal localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8668–8678, 2019. 2

[61] Yang Zhao, Zhou Zhao, Zhu Zhang, and Zhijie Lin. Cascaded prediction network via segment tree for temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4197–4206, 2021. 3

[62] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022. 2, 3, 6, 7, 12, 15

[63] 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. 2, 3

[64] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 2, 3, 6

[65] Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J Corso, and Marcus Rohrbach. Grounded video description. In *CVPR*, 2019. 2, 3

[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