# Finding "It": Weakly-Supervised Reference-Aware Visual Grounding in Instructional Videos

De-An Huang*, Shyamal Buch*, Lucio Dery, Animesh Garg, Li Fei-Fei, Juan Carlos Niebles
Stanford University
{dahuang, shyamal, ldery, garg, feifeili, jniebles}@cs.stanford.edu

## Abstract

*Grounding textual phrases in visual content with standalone image-sentence pairs is a challenging task. When we consider grounding in instructional videos, this problem becomes profoundly more complex: the latent temporal structure of instructional videos breaks independence assumptions and necessitates contextual understanding for resolving ambiguous visual-linguistic cues. Furthermore, dense annotations and video data scale mean supervised approaches are prohibitively costly. In this work, we propose to tackle this new task with a weakly-supervised framework for reference-aware visual grounding in instructional videos, where only the temporal alignment between the transcription and the video segment are available for supervision. We introduce the visually grounded action graph, a structured representation capturing the latent dependency between grounding and references in video. For optimization, we propose a new reference-aware multiple instance learning (RA-MIL) objective for weak supervision of grounding in videos. We evaluate our approach over unconstrained videos from YouCookII and RoboWatch, augmented with new reference-grounding test set annotations. We demonstrate that our jointly optimized, reference-aware approach simultaneously improves visual grounding, reference-resolution, and generalization to unseen instructional video categories.*

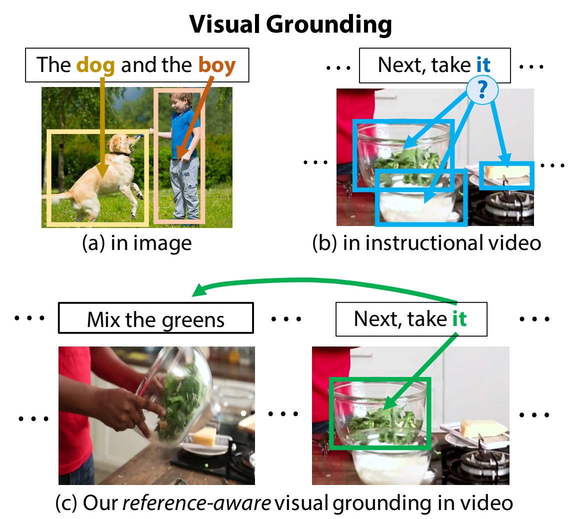


Figure 1: What is "it" in the video frame above? **(a)** Captions for visual grounding in standalone images offer fully-specified nouns or descriptors. **(b)** In contrast, instructional video captions often offer only pronouns and partially-specified descriptors, since humans can resolve the ambiguities with contextual understanding. Furthermore, structured annotations for references and groundings remain prohibitive. **(c)** To address these challenges, this work proposes a new weakly-supervised, *reference-aware* visual grounding approach that explicitly resolves the visual-linguistic meaning of referring expressions (*e.g.* "it" refers to the "greens").

## 1. Introduction

Connecting vision and language has emerged as a prominent multi-disciplinary research problem [11]. The *visual grounding* problem of connecting natural language descriptions with spatial localization in images has proved to be a critical link in solving these multi-modal tasks [19, 28, 43]. While there have been numerous studies from both natural language and vision communities that aim to address visual grounding [13, 15, 20, 25, 43, 51], both the sentences and images are obtained in a relatively controlled setting with standalone image-sentence pairs. In this work, we aim to expand this scope by studying visual grounding in instructional videos, where both the language transcription and the visual appearance are unconstrained as in real-world situations.

Visual grounding in instructional video poses two unique challenges compared to standalone image-based visual grounding: (1) Step descriptions rely heavily on pronouns and referring expressions to provide implicit links to crucial visual and linguistic context. In other words, the referring expressions (*e.g.* "it" in Fig. 1) no longer fully specify the visual appearance of entities. (2) Annotations linking the grounding and contextual references remains prohibitively

* indicates equal contribution lead author

costly in unconstrained videos. This is due to the dense nature of the graph-based annotations and the sheer scale of instructional video data [47]. While these challenges have been tackled separately, including situated language understanding in natural language processing [6, 8, 21, 26, 33] and weakly-supervised object localization [9, 13, 35, 36, 40, 46] in computer vision, simultaneously solving both for unconstrained videos remains an unsolved challenge.

To our knowledge, this is the first work to examine the challenging task of visual grounding in instructional videos. Thus, our **first contribution** is to formulate this key visual understanding task for the video domain. We introduce the *visually grounded action graph* as a structured representation to explicitly capture the latent dependencies between reference and grounding variables, and formulate grounding in videos as optimization of this graph.

Next, we address the two key technical challenges introduced by instructional video, namely context-dependent ambiguity and the prohibitive cost of labels for supervised approaches. The **second contribution** of this work is to present a novel visual grounding model that is both reference-aware and weakly-supervised. Our joint model is *reference-aware* as it explicitly resolves the situated and context-dependent meaning of referring expressions and goes beyond previous visual grounding works designed for independent image/sentence pairs. Our approach is also *weakly-supervised* in that it requires no explicit grounding supervision and only uses temporally aligned transcription and video input as supervision. The latent structure of instructional videos fundamentally breaks the independence assumption of prior standalone image-based approaches. Thus, we introduce the first reference-aware multiple instance learning (RA-MIL) framework to more effectively leverage predicted references to improve visual grounding optimization.

Because this is a new task for video understanding, our **third contribution** is to provide reference-grounding test set annotations for two main instructional video benchmarks, namely YouCookII [56] and RoboWatch [45]. We evaluate our new approach for weakly-supervised, reference-aware visual grounding in instructional videos by optimizing on over two thousand unconstrained YouTube cooking videos of the YouCookII dataset. We show that our joint approach improves grounding by explicitly modeling the latent references between sentences. We "close the loop" by further demonstrating that our learned visual grounding representations can in turn improve reference resolution within our joint framework. Finally, we demonstrate that our approach improves model generalizability to unseen instructional video categories by evaluation on RoboWatch.

## 2. Related Work

**Weakly-Supervised Localization and Visual Grounding.** Our task for visual grounding in videos builds from prior work on visual grounding with stand-alone image-sentence pairs, which aims to match entities in the caption to bounding boxes within the image. This is related to weakly-supervised object localization [9, 10, 13, 35, 36, 40, 46]. We generalize this notion to *context-dependent* referring expression localization, which adds another dimension of complexity from language understanding to our grounding problem. Recent works also aim to ground expressions in phrases beyond object categories [15, 20, 32, 34, 39, 43, 49, 54]. However, most assume the availability of ground truth annotation [15, 39, 49, 53], and all assume standalone independent image-sentence pairs [43, 18]. In this work, we jointly address the challenges from weak supervision and situated language in the instructional video domain.

**Multiple Instance Learning (MIL) in Vision.** MIL has been a effective framework for weakly-supervised learning in several applications, including image classification [50], object localization [13], tracking [5], and instance segmentation [37]. In this work, we extend the MIL approach of visual grounding in images [19] to instructional video and propose Reference-Aware MIL (RA-MIL) to effectively learn the situated referring expression in instructional video.

**Learning from Instructional Video.** In this work, we use the transcription in the instructional video for weakly-supervised visual grounding. This use of transcription as supervision has been utilized in several contexts, such as action detection [55], object state discovery [2], entity reference [16], and procedural knowledge discovery [1, 29, 45]. The most related to our work is the visual-linguistic reference resolution (VLRR) by [16], which focuses on learning entity references in the instructional video. Our work goes a step further and leverages references to solve the weakly-supervised visual grounding in instructional video.

**Reference Resolution for Visual Tasks.** We utilize reference resolution to improve visual grounding in instructional video. Recent work has used reference for improving visual tasks, such as image and 3D scene understanding [14, 24], and actor recognition [41, 44]. Here, we demonstrate that reference resolution is mutually beneficial for the challenging task of visual grounding for video understanding.

**Situated Language Understanding.** Situated language is a term in the natural language processing community capturing the notion that our own understanding of language is learned from situations and entities within them [21]. Our modeling of situated referring expression in the transcription is related to procedural text understanding in NLP [3, 6, 8, 21, 26, 30, 33]. Our work goes a step further and studies the situated language in the transcription jointly with the aligned video.

## 3. Technical Approach

Our goal is weakly-supervised visual grounding in instructional video. This is challenging since (1) the desired grounding output is latent at training, and (2) the entities

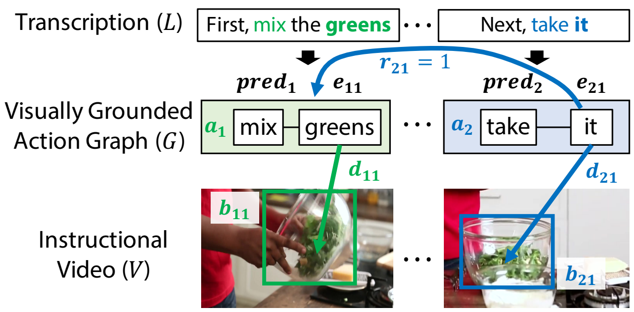


Figure 2: A visually grounded action graph ($G$) is an action graph with object box nodes $b_{ik}$ and the corresponding grounding edges $d_{ij}$ to model the visual grounding of the entities $e_{ij}$. This graph serves as the joint representation between the visual grounding within actions $a_i$ and reference resolution $r_{ij}$ between them. This reformulates visual grounding and reference resolution as finding the best set of edges $(D, R)$ in the graph given the nodes. See Section 3.1.

within the transcriptions can be highly context-dependent, with references that are also latent. We address this by formulating it as a joint optimization of a *visually grounded action graph* that explicitly captures the latent dependencies between grounding and reference (Sec. 3.1). We propose a joint framework for reference-aware visual grounding to effectively infer this graph from input video and transcription (Sec. 3.2 and 3.4). Because such dense graph annotations incur prohibitive cost in videos, we propose a new reference-aware multiple instance learning (RA-MIL) method for weakly-supervised learning (Sec. 3.3).

### 3.1. Visual Grounding Task in Videos

**Goal.** Since both the groundings and references are latent and interdependent, there is a clear need to model them in a unified manner. Inspired by the "action graph" for reference resolution in natural language text [16, 21], we propose a new *visually grounded action graph* that encompasses the latent information for both the visual grounding and reference resolution in a single explicit data structure (Figure 2). We introduce new nodes for object bounding boxes in the video frames, and new edges between each entity node and its corresponding object bounding box for visual grounding. Thus, visual grounding in instructional videos is reformulated as determining the correct "grounding edges" between entity nodes and object box nodes in the graph.

Furthermore, we seek to learn references between grounded entities and prior actions. Intuitively, this captures directed paths from starting components to final composite products. Unlike prior work [16, 21], we endeavor to learn entity-action references with the added constraint of visual grounding for entity nodes. We demonstrate that jointly resolving the latent reference is key to improving visual grounding and visual grounding also improves reference.

**Visually Grounded Action Graphs.** More formally, a visually grounded action graph $G = (E, A, B, R, D)$ has $E = \{e_{ij}\}$, a set of entity nodes $e_{ij}$ , $A = \{a_i\}$, a set of action nodes $a_i$ grouping the entity nodes with their predicates $pred_i$, $B = \{b_{ik}\}$, a set of object box nodes $b_{ik}$ aligned to each $a_i$, $R = \{r_{ij}\}$, a set of edges for the reference $r_{ij}$ of $e_{ij}$, and $D = \{d_{ij}\}$, a set of edges for the visual grounding $d_{ij}$ of $e_{ij}$. The sub-index $j$ distinguishes multiple entities within the same action $a_i$ (*e.g.* "mix *salt*, *pepper*, and *oil*"). We illustrate a portion of a graph in Figure 2. Here, each action node $a_i$ contains entity nodes $e_{ij}$, each edge $d_{ij}$ from an entity node to a object box $b_{ik}$ is a grounding, and each edge $r_{ij}$ from an entity node to an action node is a reference. Note that $G$ encompasses the information for both visual grounding ($D$) and reference resolution ($R$), where visual grounding is identical to recovering the grounding edges $D$ in the graph. Further, recovering $D$ depends on $R$, so effective visual grounding needs to be reference-aware.

**Joint Approach.** Figure 3 shows our model overview. The input is the instructional video with its time-aligned transcription, and the output is the full visually grounded action graph for the video. Graph nodes are generated by (1) parsing the transcription into entity nodes $E$ and action nodes $A$, and (2) obtaining object proposals on video frames for object box nodes $B$. In this work, we assume the nodes of the graph are provided to our joint model, and focus the task on recovering the grounding and reference edges. Such recovery is equivalent to $\text{argmax}_{D,R}\, P(D, R|E, A, B)$. We take an E-M like approach for joint optimization by alternating between optimizing the visual grounding model ($\text{argmax}_D\, P(D|E, A, B, R)$, in Section 3.2) and optimizing the reference resolution model ($\text{argmax}_R\, P(R|E, A, B, D)$, in Section 3.4).

### 3.2. Reference-Aware Visual Grounding: Model

In the previous section, we formulated reference-aware visual grounding as optimizing the grounding edges $D$ in the visually grounded action graph $G$. We now define how we parameterize our model for the probability of a grounding, $P(D|E, A, B, R)$. We decompose the full grounding model $P(D|E, A, B, R)$ into the aggregation of edge probabilities $\prod_{d\in D} P(d|E, A, B, R)$. Crucially, while instructional videos break standard independence assumptions, we can observe *conditional* independence given $E, A, B$ nodes and the references $R$ in the graph, which we also learn to infer (see Section 3.4). For $P(d|E, A, B, R)$, we model the probability of grounding an entity $e_{ij}$ to an object box $b_{lk}$. Formally, the grounding model is:

$$P(d_{ij} = (l, k)|E, A, B, R) = \text{sigmoid}(\psi(b_{lk})^T \phi_e^R(e_{ij})), \quad (1)$$

where $\phi_e^R(e_{ij})$ is a *reference-aware* entity embedding that incorporates the information of $R$ and $A$ when embedding $e_{ij}$, and $\psi(b_{lk})$ is an end-to-end trainable visual embedding.

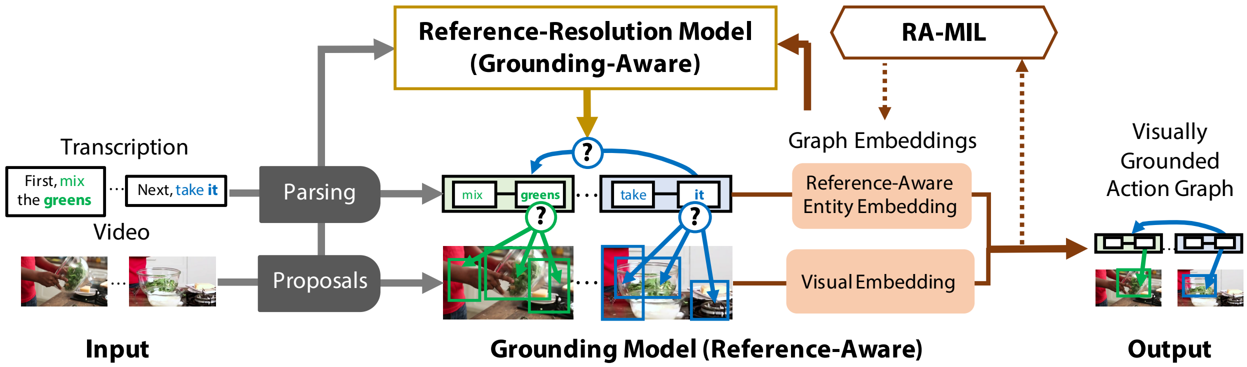


Figure 3: Overview of our model. We take as input an instructional video and its transcript, which provide us the initial entity, action, and object box nodes for the visually grounded action graph. The output of our joint model is to infer the edges of the optimal graph, including reference and grounding. We propose a grounding model that is *reference-aware*, which matches different action entities to their corresponding bounding box in the video. We design a training method for this model called reference-aware multiple instance learning (RA-MIL). Further details in Section 3.1.

Intuitively, we aim to learn the grounding model by learning a visual-semantic embedding that measures the similarity of an entity and a object box. We define these two embeddings:

**Reference-Aware Entity Embedding** $\phi_e^R(e_{ij})$. Given an entity (*e.g.* "mixture"), our goal is to embed it in a way that captures the action that it is referring to (*e.g.* "mix mayo and parsley"). We thus utilize a recursive definition for our entity embedding that is able to combine information from the referring action [16]. Thus, the entity embedding is:

$$\phi_e^R(e_{ij}) = wordEmbd(e_{ij}) + \phi_a^R(a_o), \qquad (2)$$

where $o = r_{ij}$ and $\phi_a^R(a_o) = RNN_{\theta_V}([\phi_e^R(e_{op})]_p)$ Here, $wordEmbd(\cdot)$ is the standard word embedding function (we use GloVe [38] here), $RNN_{\theta_V}$ is a recurrent neural network (RNN) embedding function [23] that takes in $[\phi_e^R(e_{op})]_p$, a list of entity embeddings of entities $e_{op}$ in action $a_o$. Here, our reference-aware entity embedding also contains the information from its referring action. This utilization of reference information in visual grounding sets our method apart from grounding models designed only for images. We show that this is important for correctly grounding entities in instructional video, where the entity is often context-dependent.

**Visual Embedding** $\psi_b(b_{lk})$. We use a deep convolutional neural network to extract the visual representation of our object boxes. In addition, an affine layer $W_V$ is added to embed the 4096-dimensional fully-connected layer representation to the dimension of the entity embedding. Formally, this can be written as $\psi_b(b_{lk}) = W_V(CNN_{\theta_V}(b_{lk}))$.

### 3.3. Reference-Aware Visual Grounding: RA-MIL

We have described the parameterization of our reference-aware visual grounding model $P(D|E, A, B, R)$. Now, we discuss the optimization objective to learn $P(D|E, A, B, R)$ with only weak supervision from temporal alignments between transcription and video segments. Inspired by recent work in visual grounding in images [13, 18], we formulate weakly-supervised visual grounding in videos as a Multiple Instance Learning (MIL) problem [4]. Herein, the supervision is provided only through the temporal alignment between the sentence and the video segment: for an entity $e_{lj}$ in step $l$, it should be grounded to one object box $b_{lk}$ from the set of all object boxes in the corresponding video segment, and there is no explicit training label for *which* box it is. The key challenge of naïvely applying an image-based framework to the video domain is that sentence-video pairs no longer follow a *strict independence assumption*. This is consequential in two key ways: (1) temporal dependence is reflected in the transcription language, which may refer to the current entity implicitly or with pronouns (*e.g.* "it"), and (2) visual grounding of the same entity is possible in multiple instruction steps with relatively high confidence, particularly in the referring actions. Because segments from the same video are heavily correlated, image-based strategies [13, 19] for negative selection can induce errors even for the labels in standard MIL approaches which assume independence.

**RA-MIL.** We address both challenges by proposing a new *Reference-Aware Multiple Instance Learning* (RA-MIL) objective to train a model to explicitly represent the dependencies between groundings caused by the references. More specifically, based on the weak supervision from the alignment (*i.e.* for step $l$, $e_{lj}$ should be grounded to $b_{lk}$ for some $k$), we first propose the following learning constraints:

$$\begin{aligned}&\max_{D_l} P(D_l|\bar{G}_l, B_l) > \max_{D_l} P(D_l|\bar{G}_l, B_m) \text{ and} \\ &\max_{D_l} P(D_l|\bar{G}_l, B_l) > \max_{D_n} P(D_n|\bar{G}_n, B_l),\end{aligned} \qquad (3)$$

for $m, n \neq l$, where $B_l = \{b_{lk}\}$ is the set of all object box nodes in the segment depicting action step $l$, and $\bar{G}_l = \{E_{1:l}, A_{1:l}, R_{1:l}\}$ be the subgraph up to segment $l$, excluding the grounding. Intuitively, the first constraint in Eq. (3) means this sub-graph $\bar{G}_l$ should have a higher probability of grounding to a box in $B_l$ in the same video segment rather than the $B_m$ of a different segment. Likewise, we have the symmetric constraint for $B_l$ given $\bar{G}_n$ of a different step.

While the model can directly utilize the reference information by operating on the subgraph $\bar{G}_l$ and can be trained with weak-supervision for reference-aware visual grounding in instructional video, we note that the constraints in Eq. (3) do not fully utilize the reference information. Consider Figure 4 as an example: while "it" is indeed grounded to the blue bounding box in the second step, it is not visually incorrect to ground it to the the bowl full of greens in the previous step, since it is the same entity. In this case, the MIL constraints in Eq. (3) are forcing the model to differentiate objects that are in fact the same with the same penalty as completely unrelated entities. Based on this intuition, we propose the following overall training loss to effectively utilize reference for weakly-supervised visual grounding:

$$\mathcal{L}_{RA-MIL} = \sum_l \Big[ \sum_m \gamma_{lm} \cdot max(0, S^R_{lm} - S^R_{ll} + \Delta) + \sum_m \gamma_{ml} \cdot max(0, S^R_{ml} - S^R_{ll} + \Delta) \Big], \tag{4}$$

where $S^R_{lm} = \sum_j \max_k \langle \phi^R_e(e_{mj}), \psi_b(b_{lk}) \rangle$ refers to the alignment score for steps $l$ and $m$ analogous to the image-sentence score in [18], and $\gamma_{lm}$ is a reference-based penalty with a value of 1.0 if step $l$ is not in the set of inferred entity-action references in step $m$. If step $l$ is present the reference set, then we set $0 < \gamma_{lm} < 1$. In this manner, the objective encourages the action graph to be grounded in the aligned video, while distinguishing penalties based on the degree to which the predicted grounding is related to the target entity.

We emphasize that RA-MIL incorporates reference-awareness in two key aspects: (1) it explicitly imposes the constraints in Eq. (3) based on the subgraph $\bar{G}_l$ to incorporate reference information of a given entity based on the relevant prior set of actions – this sets our approach apart from previous standalone image-sentence grounding methods that operate solely based on the entity expression itself [13, 19, 43]; (2) we incorporate reference-based relaxation to improve negative constraints during MIL, as per Eq. (4). We show in our experiments that both reference aspects of RA-MIL are key for visual grounding in instructional videos.

### 3.4. Grounding-Aware Reference Resolution

We have discussed our reference-aware visual grounding model $P(D|E, A, B, R)$ and our weakly-supervised training approach (RA-MIL) conditioned on the reference edges $R$. Now, we discuss how we update the contextual references given the groundings $D$ with $P(R|E, A, B, D)$, as illustrated in Figure 5. Inspired by recent frameworks using neural networks for graph optimization [17, 52], we formulate the reference edge model by proposing a hierarchical entity-action pointer network for reference resolution, based on Ptr-Net [48]. A key difference between our proposed model and a standard Ptr-Net is that we wish to link entities with prior action steps, but these exist at different hierarchical levels in the graph. Intuitively, this single-mapping

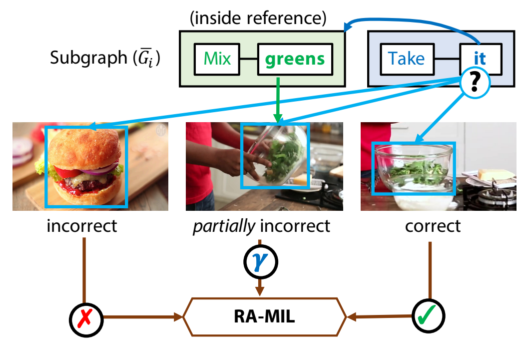


Figure 4: We propose Reference-Aware Multiple Instance Learning (RA-MIL) for reference-aware visual grounding in instructional videos by weak supervision. RA-MIL goes beyond standard MIL by (1) grounding the subgraph $\bar{G}_i$ to resolve ambiguity of situated referring expressions (*e.g.* "it" means "greens"), and (2) reference-based negative selection during MIL (*e.g.* grounding "it" to the earlier greens bounding box is not as penalized as grounding to the burger).

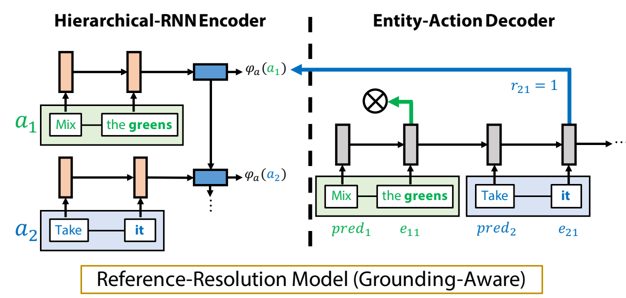


Figure 5: We propose an entity-action pointer network for reference resolution (Sec. 3.4). A hierarchical RNN encodes *action* nodes from language components. Later, we decode grounding-aware *entity* embeddings, where the output "points" to the referring action, if present.

formulation for reference resolution [21] captures the notion that some entities are causally-linked direct outputs of prior steps, where full dependency chains are obtained by traversal. Thus, we first encode the actions $a_i$ as action embeddings $\phi_a(a_i)$ using a hierarchical RNN [27]. Reference resolution occurs during decoding by a content-based attention mechanism: an RNN encodes the entity embeddings $\phi^D_e(e_{ij})$ into hidden state vectors $h^d_{ij}$, which are used to "point" back to the encoder's action embeddings or the "background action" ($\otimes$ in Fig. 5) if the entity has no reference. Formally, this is:

$$P(r_{ij} = o|E, A, B, D, H_{ij}) = \text{softmax}(u^o_{ij}), \tag{5}$$

where $u^o_{ij} = \phi_a(a_i)^T W_{att} h^d_{ij}$, and $H_{ij}$ represents all the previous entities that have been processed before $e_{ij}$. We rely on the RNN to capture the complex dependencies between $r_{ij}$ and $H_{ij}$. Importantly, we note that the entity embedding $\phi^D_e(e_{ij})$ here is grounding-aware as it summarizes the *visual* information in the linked object box. To this end,

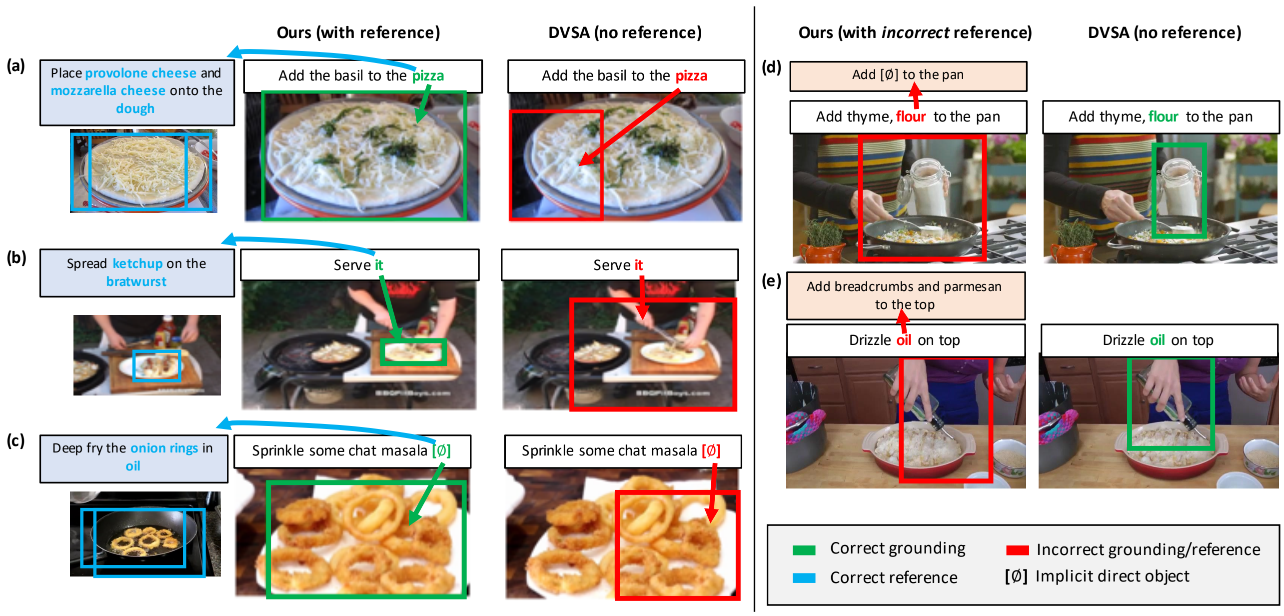


Figure 6: Qualitative results of our reference-aware visual grounding approach with RA-MIL. **(a, b, c)** Our approach improves visual grounding by explicitly resolving the meaning of ambiguous context-dependent referring expressions during optimization. We highlight improvements with **(a)** expressions that are outputs of prior steps ("pizza"), **(b)** pronouns ("it"), and **(c)** implicit direct objects (denoted as $[\emptyset]$ [16, 21]). **(d, e)** Since references are also inferred by our joint model, incorrect reference predictions can lead to lower grounding quality, compared with standalone image approaches (DVSA [18]). Note that we show *portions* of the output visually grounded action graph above, and include longer visualizations in the supplement.

we define $\phi_e^D(e_{ij}) = W_e^D[wordEmbd(e_{ij}); CNN(b_{id_{ij}})]$, where $W_e^D$ is a linear transformation to combine information from both the entity and the object box into a single embedding. We verify in our experiments that reference resolution improves grounding in a mutually beneficial manner.

### 3.5. Learning & Inference

**Visual Grounding.** As the objective for RA-MIL is fully differentiable, we are able to use backpropagation to optimize the full reference-aware visual grounding model with weak-supervision. Once the reference-aware grounding edge model in Section 3.2 is trained, the inference for $\operatorname{argmax}_D P(D|E, A, B, R)$ is a greedy score maximization in the aligned action, since we assume conditional independence between grounding edges given inferred references.

**Reference Resolution.** We follow the hard-EM approach in [16] for reference resolution. We apply a cross entropy classification loss over the decoding output in Eq. (5), comparing against the current best estimated graph. Inference can be a single forward pass of our reference resolution model. We initialize the reference edges $R$ by unsupervised reference resolution from [16]. We alternate training our grounding and reference models after initialization.

## 4. Experiments

Given a referring expression such as "mixture" in the instructional video, our goal is to visually ground it to the corresponding object bounding box in the video, while also resolving its contextual reference. In this section, we discuss our experiments to evaluate our joint approach for grounding, reference resolution, and generalizability.[1]

**Dataset and Annotation.** For weakly-supervised training, we use the YouCookII dataset [56], which is a large-scale dataset of over 2000 *unconstrained* instructional videos from 90 cooking recipes from YouTube. Each video recipe contains 3 to 15 steps (*i.e.* actions in our graph), where each step description is a temporally-aligned imperative sentence provided by the dataset. Because we are proposing a new task, for *evaluation* we provide new annotations for reference-grounding for a subset containing representative videos. Annotations and procedure details are provided in our supplementary, as well as discussion of automatic speech recognition (ASR) output as a potential source of instructional transcription input. We emphasize that *none* of this new information is used during training for our reference-aware visual grounding model for our main experiments.

Furthermore, for our generalizability analysis, we leverage the test set of the RoboWatch dataset [45], which contains instructional videos annotated with groundtruth temporal intervals and step captions. Once again, we annotate extra ground truth information for reference and grounding in each video. In total, we provide over 15 hours of video with dense entity-action node, reference, and grounding annotations

[1]Please refer to our project website for supplementary material.

Table 1: Weakly supervised visual grounding results (Top-1 accuracy) on YouCookII. We observe improvement in visual grounding across simple, medium, and hard graph complexity subsets with our method. See Section 4.1 for details.

| Method | YC-S | YC-M | YC-H | YC-All |
|---|---|---|---|---|
| Proposal Upper Bnd. | 67.4% | 65.1% | 64.1% | 65.5% |
| Random | 6.5% | 10.2% | 8.7% | 8.4% |
| DVSA [18] | 17.9% | 22.5% | 18.2% | 20.7% |
| Ours w/o Relaxation | 26.6% | 25.5% | 23.6% | 25.2% |
| Ours Full (RA-MIL) | **28.6%** | **27.7%** | **24.0%** | **26.7%** |

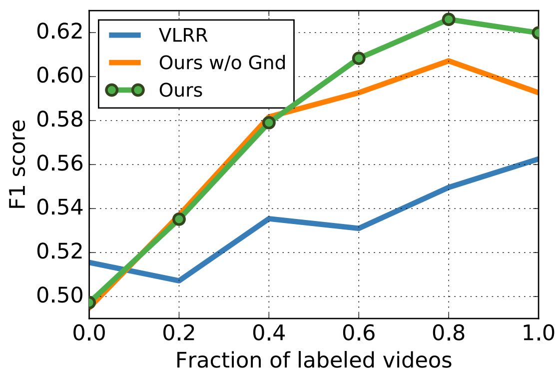


Figure 7: Reference resolution results (Sec. 4.2) on YouCookII subset. Our proposed entity-action pointer network model outperforms the VLRR [16] baseline, and we observe visual grounding can improve reference resolution.

across 2 distinct instructional video datasets.

**Implementation Details.** We parse the step description by the Stanford CoreNLP parser [31] into actions and entities. For each video, we subsample five frames per video segment for both training and testing. For each frame, we use the RPN from Faster R-CNN [42] for proposing the object boxes in the frames. For comparison to prior work [18], we use the top-20 proposal detections in a frame. Since YouCookII does not have parsed entity/action annotations, we leverage automatic parsing for training only, and provide corrected entity and action nodes as input during inference. We use Adam [22] for optimization and a learning rate 0.001. We clip gradients elementwise at 5 and use 0.3 dropout for regularization. Additional implementation details are included as part of supplementary material.

## 4.1. Evaluating Visual Grounding

**Experimental Setup.** First, we learn our model by optimizing on all the instructional videos in the YouCookII dataset [56] with only weak supervision from transcription-video temporal alignment. Parsed action $A$, entity $E$ and generated object box $B$ nodes are provided as input, as per Section 3.1. Inference on reference resolution and visual grounding follows Section 3.5. We follow prior work [12, 43] and compute accuracy as the ratio of phrases for which the grounded bounding box overlaps with the ground-truth by more than 0.5 Intersection-over-Union (IoU).

Table 2: Generalizability to *unseen* instructional video classes (RoboWatch). We observe stronger generalization performance with our reference-aware visual grounding method. See Section 4.3 for details.

| Method | RW-Cook | RW-Misc | RW-All |
|---|---|---|---|
| Proposal Upper Bnd. | 63.0% | 48.4% | 56.3% |
| Random | 10.4% | 6.2% | 9.0% |
| DVSA [18] | 22.4% | 12.6% | 17.5% |
| Ours w/o Relaxation | 23.8% | 10.4% | 18.0% |
| Ours Full (RA-MIL) | **26.8%** | **14.3%** | **19.8%** |

**Grounding Approaches.** We compare to the following models and variations of our model for visual grounding:

- *Deep Visual-Semantic Alignment (DVSA)* [18]. This is a standard weakly-supervised image-based visual grounding method without the reference information, which leverages standard multiple-instance learning. Notably, we compare to this standalone image approach since it can most directly be considered an ablation of our method without reference.
- *Ours w/o Relaxation.* This method uses the loss in Eq. (4), but does *not* utilize the reference information in negative selection ($\gamma$). Importantly, it still grounds the full subgraph $\bar{G}_l$, which means it does incorporate reference information. This baseline is an ablation of our method indicating the need for both reference-aware aspects of RA-MIL.
- *Our full approach (RA-MIL).* This is our full joint model leveraging the full RA-MIL formulation, as in Section 3.

**Limitations.** Since grounding is highly dependent on the input bounding box nodes, we also report the upper bound performance if the *best* matching proposals were chosen by some method. We observe that this is approximately 65%, which is less than upper bounds of 78% reported on standalone image datasets for visual grounding like Flickr30K [43] and may reflect difficulties introduced by noisy images in unconstrained instructional video. We discuss additional limitations due to the multiple-instance learning paradigm and parsing errors during training in the supplementary.

**Results.** The results of these weakly-supervised visual grounding models on YouCookII are shown in Table 1. Our full method outperforms the baseline and ablation methods, including DVSA [18] which is not reference-aware. We observe that grounding the subgraph $\bar{G}_l$ containing the reference information to resolve the meaning of referring expressions, rather than the raw entity itself is important. Qualitative results are shown in Figure 6. We observe the resolved meaning of the referring expression indeed improves the grounding performance, though overall it remains limited by constraints of weak supervision and dependency on input bounding boxes. By grounding $\bar{G}_l$, RA-MIL links referring actions with the visual appearance of the entity in the current and contextual frames. We include longer-form graph visualizations and additional discussion in our supplementary. While reference can help visual grounding in the

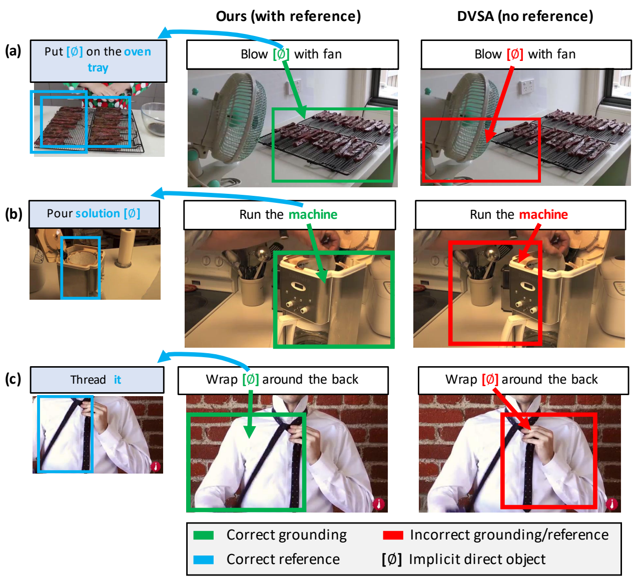


Figure 8: Qualitative results for generalization experiments, as described in Section 4.3. We evaluate our model trained on cooking tasks from YouCookII on **(a)** unseen recipes (*e.g.* making beef jerky), and **(b, c)** unseen instructional video categories (*e.g.* cleaning a coffee machine, tying a tie).

instructional video, incorrect reference predictions can lead to incorrect grounding predictions as shown in Figure 6(d,e).

### 4.2. Visual Grounding for Reference Resolution

In this section, we examine both (1) the proposed entity-action pointer network as a reference resolution architecture, and (2) the impact of grounding on reference resolution.

For this self-contained experiment, we compare against the prior Visual-Linguistic Reference Resolution approach (VLRR) in [16], and report the F1 measure as defined in [21] over different supervision levels. We benchmark performance on a *subset* of YouCookII, performing multiple 2:1 train-test splits of the 90 recipes and varying the ratio of the provided graphs for training. Full experiment details and discussion of grounding impact during our weakly-supervised reference training is included in the supplement. The results are shown in Figure 7. Here, ratio 0.0 means no input graphs are used for training, and ratio 1.0 means that all 60 training graphs are used. Understandably, the unsupervised VLRR baseline has slightly higher performance with no labels in the training set. This is likely due to strong constraints inherent to the unsupervised VLRR model design, which are not present in our weakly-supervised pointer network architecture. However, we observe that our entity-action pointer network quickly outperforms the VLRR baseline even with a few additional labels. Furthermore, as the training set increases to sufficient size, visual grounding ultimately proves effective for improving reference resolution. We emphasize that the overall number of graphs at ratio 1.0 is still far smaller than the overall training set, which is used in the main reference-aware visual grounding experiments.

### 4.3. Generalizability

We further evaluate the ability of our model to *generalize* to unseen classes of instructional video in the RoboWatch dataset [45], which includes 20 classes that each correspond to a top "How to" web query. We draw inspiration from prior work in action localization [7] for our experiment design. Here, we train the models on YouCookII as before, but run inference on the RoboWatch test set, augmented with new reference and visual grounding groundtruth annotations. We also examine performance on subsets with cooking-specific (containing unseen recipes) and miscellaneous videos, which includes classes such as "How to Unclog Bathroom Drain" and "How to Clean a Coffee Maker". In all cases, we ensure that there is no recipe or video overlap with YouCookII.

We report generalization performance in Table 2, and include qualitative visualizations in Figure 8. We observe that our full approach with RA-MIL outperforms the other methods at generalization. For cooking-specific videos, we observe stronger generalization to visual grounding for unseen recipes. Interestingly, we also show some improved generalization to the "Misc" subset as well, despite the domain gap between the cooking videos in YouCookII and the other instruction categories present here. The decrease in the proposal upper bound for miscellaneous tasks indicates that generalizability of these models is also limited by the visual encoder and proposals method. This suggests that improving proposals, particularly for the noisy images present in unconstrained videos, may be critical for general application of this technique for practical purposes.

## 5. Conclusion

We propose a new reference-aware approach for weakly-supervised visual grounding in instructional video. We introduce the visually grounded action graph and formulate the task as optimization for both reference and grounding edges. Our proposed Reference-Aware MIL (RA-MIL) effectively leverages references for visual grounding in a unified framework. We provide new annotations over two main instructional video datasets for visually-grounded action graphs. Our experiments verify that resolving the meaning of situated and context-dependent referring expression is important for visual grounding in instructional video, and that visual grounding can further improve reference resolution. Finally, we show that our joint reference-aware approach improves generalizability to unseen instructional video categories.

**Acknowledgements.** This research was sponsored in part by grants from Toyota Research Institute (TRI) and the Office of Naval Research (N00014-15-1-2813). This article reflects the authors' opinions and conclusions, and not of any Toyota entity. We thank our anon. reviewers, L. Zhou, O. Sener, S. Yeung, J. Ji, and J. Emmons for their helpful comments and discussion.

## References


[1] J.-B. Alayrac, P. Bojanowski, N. Agrawal, I. Laptev, J. Sivic, and S. Lacoste-Julien. Unsupervised learning from narrated instruction videos. In *CVPR*, 2016. 2

[2] J.-B. Alayrac, J. Sivic, I. Laptev, and S. Lacoste-Julien. Joint discovery of object states and manipulating actions. *arXiv:1702.02738*, 2017. 2

[3] J. Andreas and D. Klein. Alignment-based compositional semantics for instruction following. In *EMNLP*, 2015. 2

[4] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. *NIPS*, 2003. 4

[5] B. Babenko, M.-H. Yang, and S. Belongie. Robust object tracking with online multiple instance learning. *TPAMI*, 33(8):1619–1632, 2011. 2

[6] A. Bordes, N. Usunier, R. Collobert, and J. Weston. Towards understanding situated natural language. In *AISTATS*, 2010. 2

[7] S. Buch, V. Escorcia, C. Shen, B. Ghanem, and J. C. Niebles. SST: Single-stream temporal action proposals. In *CVPR*, 2017. 8

[8] D. L. Chen and R. J. Mooney. Learning to interpret natural language navigation instructions from observations. In *AAAI*, 2011. 2

[9] T. Deselaers, B. Alexe, and V. Ferrari. Weakly supervised localization and learning with generic knowledge. *International journal of computer vision*, 100(3):275–293, 2012. 2

[10] S. Divvala, A. Farhadi, and C. Guestrin. Learning everything about anything: Webly-supervised visual concept learning. 2014. 2

[11] F. Ferraro, N. Mostafazadeh, T.-H. K. Huang, L. Vanderwende, J. Devlin, M. Galley, and M. Mitchell. A survey of current datasets for vision and language research. 2015. 1

[12] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. 2016. 7

[13] R. Gokberk Cinbis, J. Verbeek, and C. Schmid. Multi-fold mil training for weakly supervised object localization. In *CVPR*, 2014. 1, 2, 4, 5

[14] M. Hodosh, P. Young, C. Rashtchian, and J. Hockenmaier. Cross-caption coreference resolution for automatic image understanding. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, 2010. 2

[15] R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, and T. Darrell. Natural language object retrieval. In *CVPR*, 2016. 1, 2

[16] D.-A. Huang, J. J. Lim, J. C. Niebles, and L. Fei-Fei. Unsupervised visual-linguistic reference resolution in instructional videos. In *CVPR*, 2017. 2, 3, 4, 6, 7, 8

[17] D. D. Johnson. Learning graphical state transition. In *ICLR*, 2017. 5

[18] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 2, 4, 5, 6, 7

[19] A. Karpathy, A. Joulin, and L. Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *NIPS*, 2014. 1, 2, 4, 5

[20] S. Kazemzadeh, V. Ordonez, M. Matten, and T. L. Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014. 1, 2

[21] C. Kiddon, G. T. Ponnuraj, L. Zettlemoyer, and Y. Choi. Mise en place: Unsupervised interpretation of instructional recipes. In *EMNLP*, 2015. 2, 3, 5, 6, 8

[22] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[23] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302, 2015. 4

[24] C. Kong, D. Lin, M. Bansal, R. Urtasun, and S. Fidler. What are you talking about? text-to-image coreference. In *CVPR*, 2014. 2

[25] J. Krishnamurthy and T. Kollar. Jointly learning to parse and perceive: Connecting natural language to the physical world. *TACL*, 1:193–206, 2013. 1

[26] T. A. Lau, C. Drews, and J. Nichols. Interpreting written how-to instructions. In *IJCAI*, 2009. 2

[27] J. Li, M.-T. Luong, and D. Jurafsky. A hierarchical neural autoencoder for paragraphs and documents. In *ACL*, 2015. 5

[28] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In *NIPS*, 2016. 1

[29] J. Malmaud, J. Huang, V. Rathod, N. Johnston, A. Rabinovich, and K. Murphy. What's cookin'? interpreting cooking videos using text, speech and vision. In *NAACL HLT*, 2015. 2

[30] J. Malmaud, E. J. Wagner, N. Chang, and K. Murphy. Cooking with semantics. In *Proceedings of the ACL 2014 Workshop on Semantic Parsing*, 2014. 2

[31] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky. The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*, pages 55–60, 2014. 7

[32] J. Mao, J. Huang, A. Toshev, O. Camburu, A. Yuille, and K. Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016. 2

[33] D. K. Misra, K. Tao, P. Liang, and A. Saxena. Environment-driven lexicon induction for high-level instructions. In *ACL*, 2015. 2

[34] V. K. Nagaraja, V. I. Morariu, and L. S. Davis. Modeling context between objects for referring expression understanding. In *ECCV*, 2016. 2

[35] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *CVPR*, 2015. 2

[36] M. Pandey and S. Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. In *ICCV*, 2011. 2

[37] D. Pathak, E. Shelhamer, J. Long, and T. Darrell. Fully convolutional multi-class multiple instance learning. In *ICLR Workshop*, 2015. 2

[38] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014. 4

[39] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015. 2

[40] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari. Learning object class detectors from weakly annotated video. In *CVPR*, 2012. 2

[41] V. Ramanathan, A. Joulin, P. Liang, and L. Fei-Fei. Linking people in videos with "their" names using coreference resolution. In *ECCV*, 2014. 2

[42] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 7

[43] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele. Grounding of textual phrases in images by reconstruction. In *ECCV*, 2016. 1, 2, 5, 7

[44] A. Rohrbach, M. Rohrbach, S. Tang, S. J. Oh, and B. Schiele. Generating descriptions with grounded and co-referenced people. In *CVPR*, 2017. 2

[45] O. Sener, A. R. Zamir, S. Savarese, and A. Saxena. Unsupervised semantic parsing of video collections. In *ICCV*, 2015. 2, 6, 8

[46] H. O. Song, R. Girshick, S. Jegelka, J. Mairal, Z. Harchaoui, and T. Darrell. On learning to localize objects with minimal supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2014. 2

[47] K. Tang, V. Ramanathan, L. Fei-Fei, and D. Koller. Shifting weights: Adapting object detectors from image to video. In *NIPS*, 2012. 2

[48] O. Vinyals, M. Fortunato, and N. Jaitly. Pointer networks. In *NIPS*, 2015. 5

[49] L. Wang, Y. Li, and S. Lazebnik. Learning deep structure-preserving image-text embeddings. In *CVPR*, 2016. 2

[50] J. Wu, Y. Yu, C. Huang, and K. Yu. Deep multiple instance learning for image classification and auto-annotation. In *CVPR*, 2015. 2

[51] F. Xiao, L. Sigal, and Y. J. Lee. Weakly-supervised visual grounding of phrases with linguistic structures. In *CVPR*, 2016. 1

[52] D. Xu, Y. Zhu, C. Choy, and L. Fei-Fei. Scene graph generation by iterative message passing. In *CVPR*, 2017. 5

[53] S. Yang, Q. Gao, C. Liu, C. Xiong, S.-C. Zhu, and J. Y. Chai. Grounded semantic role labeling. In *Proceedings of NAACL-HLT*, 2016. 2

[54] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg. Modeling context in referring expressions. In *ECCV*, 2016. 2

[55] S.-I. Yu, L. Jiang, and A. Hauptmann. Instructional videos for unsupervised harvesting and learning of action examples. In *ACM MM*, 2014. 2

[56] L. Zhou, C. Xu, and J. J. Corso. Procnets: Learning to segment procedures in untrimmed and unconstrained videos. *arXiv:1703.09788*, 2017. 2, 6, 7

# Finding “It”: Weakly-Supervised Reference-Aware Visual Grounding in Instructional Video

## (Supplementary Material)

## A1. Supplementary Material Overview

In this document, we provide supporting analysis and additional details for our main paper (see `finding-it.github.io` for more):

- **Additional Visualizations:** We provide additional discussion of errors inherent to the limitations of weak supervision. For video and extended visualizations please refer to the oral presentation video.

- **Experimental Details:** We include additional experimental and dataset details. Note that our new annotations for YouCookII and RoboWatch are available on the project page, to compare against and build upon our work.

- **Performance Breakdown:** We provide additional results and analysis, including a more detailed breakdown of the performance of our reference-aware method at different Top-K and IoU thresholds.

## A2. Additional Visualizations

**Graph and Video Visualization.** Please refer to our oral presentation video for examples, including a walkthrough of a full visually-grounded action graph on a video describing how to prepare a spaghetti and meatballs dish.

**Error Analysis.** In the main paper, we included negative examples where incorrectly inferred reference edges negatively impacted the grounding performance *relative* to when no reference is incorporated. In Figure A1, we include additional visualizations of other classes of errors, such as those *inherent* to the weak supervision method from our work (based on multiple instance learning).

We observe that scene clutter, number of co-occuring entities, and size of the entities can affect model performance. This is a common limitation arising from the multiple instance learning objective, since the only supervision that is provided during training is based on the transcription-video alignment (*i.e*. grounding in the same *segment* is encouraged). Notably, such errors are not specific to entities that contain references to prior steps. Note that no explicit supervision is provided *within* the segment for *which* specific bounding box contains the entity. This knowledge is weakly supervised through the implicit overlap of multiple different segments, which may contain the same entity.

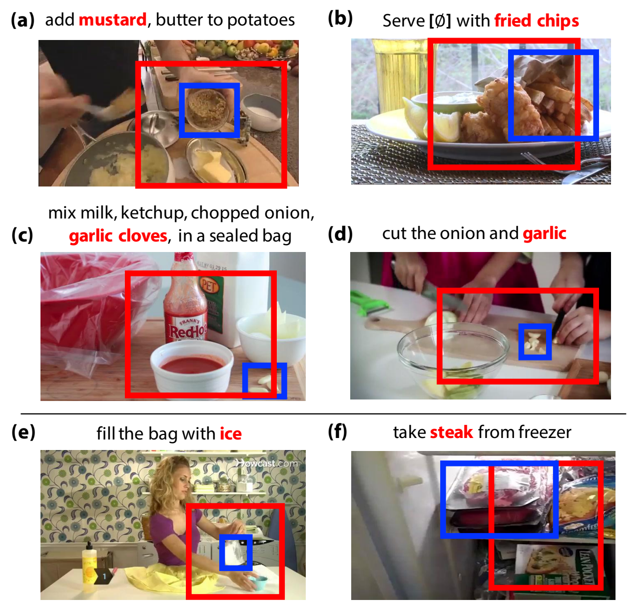


Figure A1: Additional negative results for **(a-d)** YouCookII and **(e-f)** RoboWatch. Red bounding box indicates incorrect grounding of the entity in red/bold font. Blue bounding box shows corresponding groundtruth box. We observe errors inherent to weak supervision, where no direct supervision is provided over entity localization. See Section A2 for additional **error analysis** discussion.

This naturally poses a problem when entities predominantly co-occur across segments, in multiple aligned transcription-video segment pairs. We can see this in Figure A1(c-d), where onion and garlic often occur in the same step in different videos, so the model has difficulty distinguishing the two. The other ingredients in (c) have better grounding (not shown in the figure), which may be due to entities like

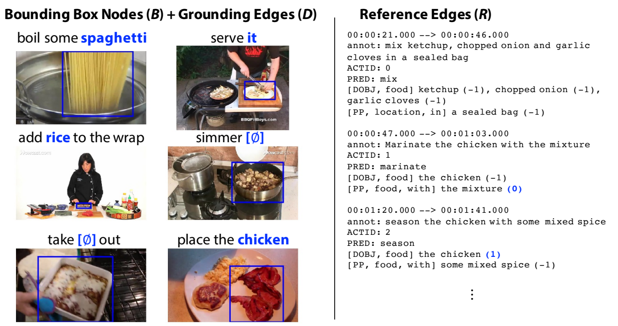


Figure A2: Example annoations for bounding box nodes ($B$), grounding edges ($G$), and reference edges ($R$) that we provide for evaluation. **(left)** Note that since this is video data, there are many "groundtruth" boxes (across frames) that correspond to each grounded entity. Further, each frame may contain multiple entities - we only show 1 frame and 1 entity for each example above. **(right)** we show an excerpt of our reference annotations file, showing how outputs from steps 0 and 1 are referred in later steps. See Section A3.

Table A1: Top-1 performance of our method at different Intersection over Union (IoU) thresholds. We observe improvement with our reference-aware approach for low, medium, and high IoU thresholds. See Section A4 for details.

| **Method** | **IoU=0.3** | **IoU=0.5** | **IoU=0.7** |
|---|---|---|---|
| Proposal Upper Bnd. | 75.8% | 65.5% | 44.5% |
| Random | 23.5% | 8.4% | 1.6% |
| DVSA [17] | 40.2% | 20.7% | 5.2% |
| Ours Full (RA-MIL) | **43.7%** | **26.7%** | **8.1%** |

milk and ketchup occurring separately in other segment pairs during training. Localizing small entities and visually similar entities in cluttered scenes, shown in Figure A1(a-b) can prove challenging by similar reasoning. We observe similar errors during generalization experiments as well, as shown in Figure A1(e-f).

We suspect that application of this work for high performance video understanding systems may require finer grained step parsing to reduce co-occuring expressions in the same time segment, as well as some degree of training set annotation (perhaps incorporating semi-supervision) to overcome such errors.

## A3. Experimental Details

In this section we describe additional details for our experiments and evaluation protocols. See Figure A2 for example annotations.

**Recurrent Neural Network.** For our recurrent neural network (described as RNN in our Technical Approach), we leverage a bidirectional long-short term memory (LSTM) network. Such bidirectional networks encode the input sequence over a forward and backward pass, and have been demonstrated to have slightly higher performance on natural language and speech tasks over single-pass recurrent networks [17].

Table A2: Top-$K$ performance of our method at different values of $K$ with IoU=0.5. We observe consistent improvement at multiple thresholds. See Section A4 for details.

| **Method** | $K=1$ | $K=3$ | $K=5$ |
|---|---|---|---|
| Proposal Upper Bnd. | 65.5% | 65.5% | 65.5% |
| Random | 8.4% | 19.9% | 27.9% |
| DVSA [17] | 20.7% | 31.8% | 38.0% |
| Ours Full (RA-MIL) | **26.7%** | **37.1%** | **42.5%** |

**Evaluation subsets.** To better understand the evaluation performance of our reference-aware visual grounding method, we considered different subsets of the YouCookII and RoboWatch test sets. We proposed and evaluated our approach on three mutually-exclusive subsets (YC-S, YC-M, YC-H) of YouCookII based on their graph complexity. Simple graph complexity means videos that have graph nodes with low in/out-degree and are relatively short in duration. Videos with groundtruth graphs with higher degree nodes and are longer would be categorized to YC-M or YC-H. For RoboWatch, since our purpose was to evaluate generalizability, we ensured that there was no video or category overlap between this test set and YouCookII, which was used as the training set. We then also considered two subsets, one focused on *unseen* recipes (RW-Cook) and the other focused on unseen instruction categories (RW-Misc).

**Automatic parsing.** The focus of our work was on developing a method that would take input nodes (after parsing and object proposals) and infer the optimal reference and grounding edges in the output visually grounded action graph. Thus, our approach is in many ways bottlenecked by the quality of the input proposals and parsed entities. During training, we leverage the Stanford CoreNLP parser [30] to automatically parse entities. However, since this parser is fine-tuned for newspaper datasets (frequent in natural language processing), we added a few hard-coded rules to improve predicate and prepositional phrase parsing. Nonetheless, there is still some noise introduced by such parsing. We believe that as further progress is made in the NLP community for parsing algorithms and toolkits, our method may benefit from reduced training noise. Note that during evaluation, we have correct parsed entities - improvements in automatic parsing would only improve the *training* aspect of our approach.

## A4. Performance Breakdown

As part of our additional results, we consider a more detailed performance breakdown of our method along other standard metrics to give a more comprehensive evaluation.

**IoU Performance.** We consider the performance of our method at different thresholds of intersection-over-union (IoU), which is a metric for how much the grounded bounding box overlaps with the groundtruth. In the main paper, we report results at a fixed threshold of 0.5, as is standard practice. In Table A1, we report results at low (0.3) and high (0.7) thresholds as well. We observe that our joint reference-aware approach with RA-MIL provides higher performance across the range, with higher relative gains at higher localization thresholds.

Interestingly, we also observe that the relative perfomance increase of $\sim 20\%$ roughly corresponds to the overall fraction of ambiguous entities (*e.g*. "it", implicit direct objects, etc.) in the YouCookII dataset. This is line with qualitative observations indicating improved grounding of such ambiguous terms with our reference-aware appraoch.

**Top-$K$ Performance** We also consider the performance of our method at different values of $K$. In the main paper, we report all results with Top-1 accuracy, which means we only consider the top 1 ranked grounded bounding box. For completeness, we include results at Top-3 and Top-5 values as well in Table A2. Naturally, we observe the greatest improvement in relative performance at stricter thresholds of K. However, we do find consistent improvement even at higher threshold values.