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

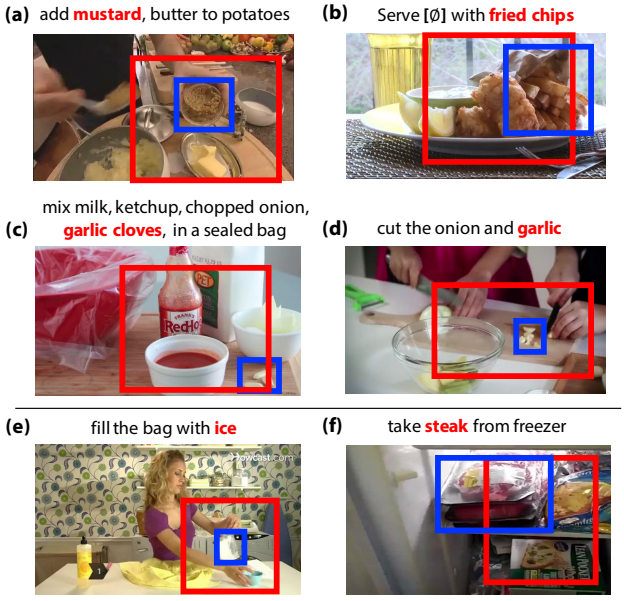

Figure A1: Additional negative results for **(a-d)** YouCookII and **(e-f)** RoboWatch. Red bounding box indicates incorrect grounding of the entity in red/bold font. Blue bounding box shows corresponding groundtruth box. We observe errors inherent to weak supervision, where no direct supervision is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

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.

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

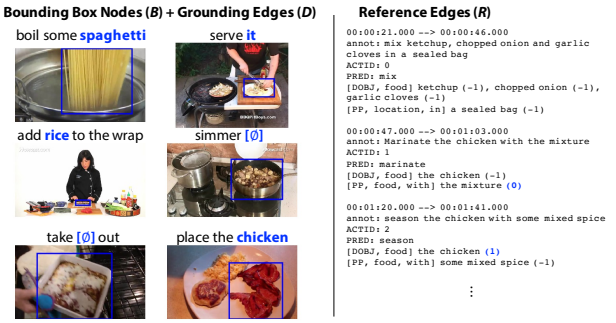

Figure A2: Example annoations for bounding box nodes (B), grounding edges (G), and reference edges (R) that we provide for evaluation. **(left)** Note that since this is video data, there are many "groundtruth" boxes (across frames) that correspond to each grounded entity. Further, each frame may contain 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.