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.

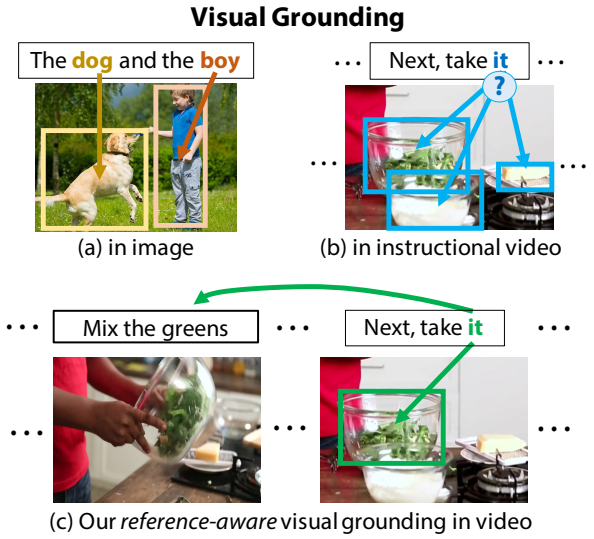

Figure 1: What is "it" in the video frame above? **(a)** Captions for visual grounding in standalone images offer fully-specified nouns or descriptors. **(b)** In contrast, instructional video captions often offer only pronouns and partially-specified descriptors, since humans can resolve the ambiguities 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").

References

[1] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 2

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

* indicates equal contribution lead author

(Note: Please refer to main CVPR2018 proceedings for full paper.)

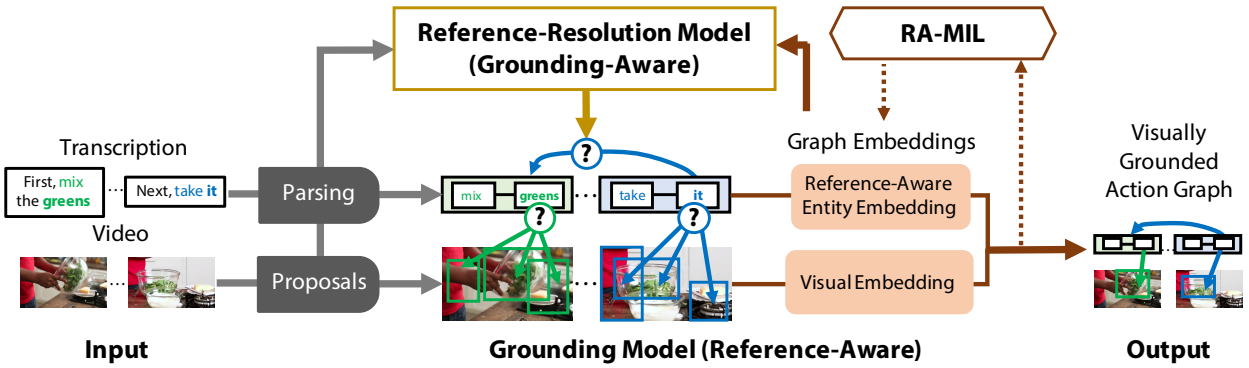

Figure 2: Overview of our model. We take as input an instructional video and its transcript, which provide us the initial entity, action, and object box nodes for the visually grounded action graph. The output of our joint model is to infer the edges of the optimal graph, including reference and grounding. 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).

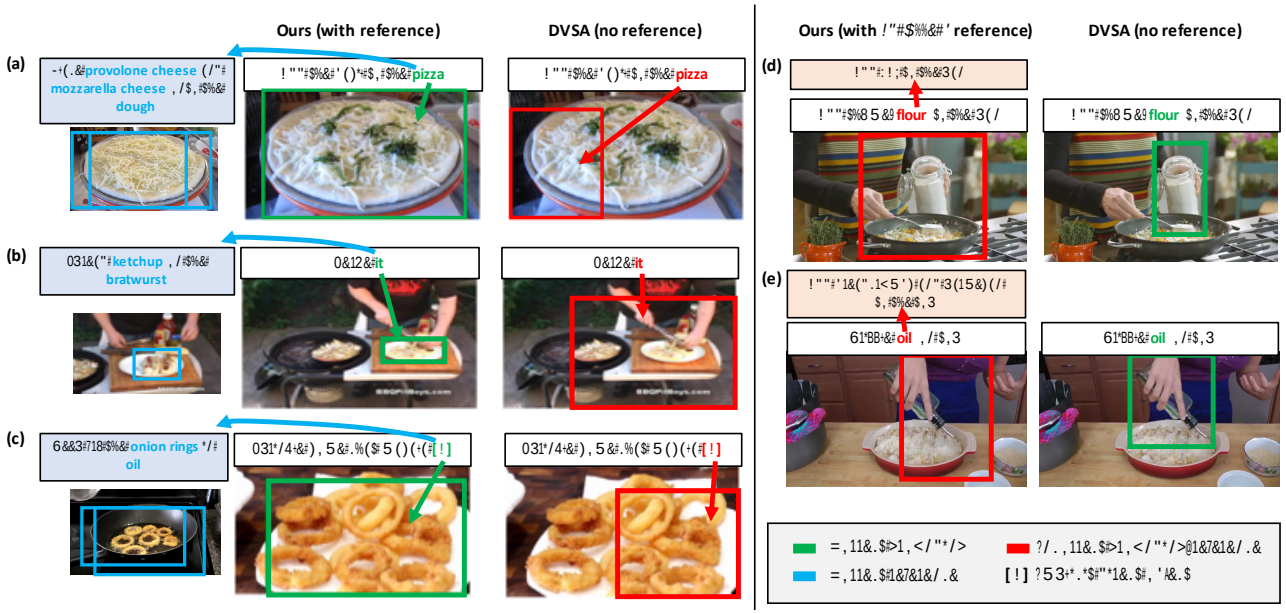

Figure 3: Qualitative results of our reference-aware visual grounding approach with RA-MIL. **(a, b, c)** Our approach improves visual grounding by explicitly resolving the meaning of ambiguous context-dependent referring expressions during optimization. We highlight improvements with **(a)** expressions that are outputs of prior steps ("pizza"), **(b)** pronouns ("it"), and **(c)** implicit direct objects (denoted as [;] [2]). **(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 [1]). Note that we show *portions* of the output visually grounded action graph above, and include longer visualizations in the supplement.