Supplementary Material: Resolving Language and Vision Ambiguities Together: Joint Segmentation & Prepositional Attachment Resolution in Captioned Scenes

Gordon Christie[1,*], **Ankit Laddha**[2,*], **Aishwarya Agrawal**[1], **Stanislaw Antol**[1]
Yash Goyal[1], **Kevin Kochersberger**[1], **Dhruv Batra**[3,1]
[1]Virginia Tech [2]Carnegie Mellon University [3]Georgia Institute of Technology
ankit1991laddha@gmail.com
{gordonac,aish,santol,ygoyal,kbk,dbatra}@vt.edu

Abstract

In this supplementary material, we provide the following:

1: Additional motivation for our MEDIATOR model.
2: Background on ABSTRACT-50S.
3: Details of the dataset curation process for the ABSTRACT-50S, PASCAL-50S, and PASCAL-Context-50S datasets.
4: Results where we study the effect of varying the weighting of each module in our approach.
5: Performances and gains over the independent baseline on PASCAL-Context-50S for each preposition.
6: Qualitative examples from our approach.

1 Additional Motivation for MEDIATOR

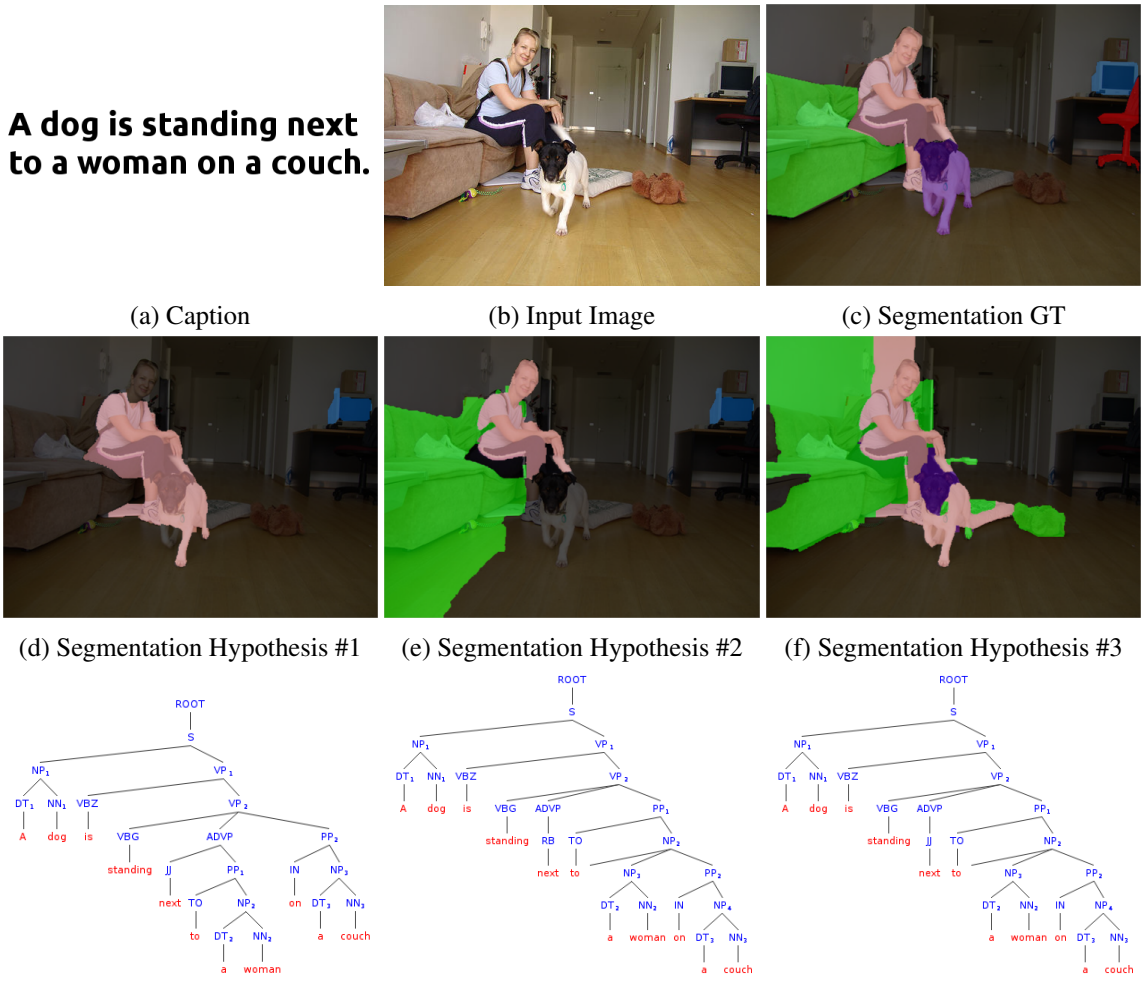
An example providing additional motivation for our approach is shown in Figure 1, where the ambiguous sentence that describes the image is "A dog is standing next to a woman on a couch". The ambiguity is "(dog next to woman) on couch" vs "dog next to (woman on couch)", which is reflected in parse trees' uncertainty. Parse tree #1 (Figure 1g) shows "standing" (the verb phrase of the noun "dog") connected with "couch" via the "on" preposition, whereas parse trees #2 (Figure 1h) and #3 (Figure 1i) show "woman" connected with "couch" via the "on" preposition. This ambiguity can be resolved if we look at an accurate semantic segmentation such as Hypothesis #3 (Figure 1f) of the associated image (Figure 1b). Likewise, we might be able to do better at semantic segmentation if we choose a segmentation that is consistent with the sentence, such as Segmentation Hypothesis #3 (Figure 1f), which contains a person on a couch with a dog next to them, unlike the other two hypotheses (Figure 1d and Figure 1e).

2 Background About ABSTRACT-50S

The Abstract Scenes dataset (Zitnick and Parikh, 2013) contains synthetic images generated by human subjects via a drag-and-drop clipart interface. The subjects are given access to a (random) subset of 56 clipart objects that can be found in park scenes, as well as two characters, Mike and Jenny, with a variety of poses and expressions. Example scenes can be found in Figure 2. The motivation is to allow researchers to focus on higher-level semantic understanding without having to deal with noisy information extraction from real images, since the entire contents of the scene are known exactly, while also providing a dense semantic space to study (due to the heavily constrained world). We used the dataset in precisely this way to first test out the PPAR module in isolation to demonstrate that this problem can be helped by a sentence's corresponding image.

3 Dataset Curation and Annotation

The subsets of the PASCAL-50S and ABSTRACT-50S datasets used in the main paper were carefully curated by two vision + NLP graduate students. The subset of the PASCAL-Context-50S dataset used in the main paper was curated by Amazon Mechanical Turk (AMT) workers. The following describes the dataset curation process for each datset.



(a) Caption (b) Input Image (c) Segmentation GT

(d) Segmentation Hypothesis #1 (e) Segmentation Hypothesis #2 (f) Segmentation Hypothesis #3

(g) Parse Hypothesis #1 (h) Parse Hypothesis #2 (i) Parse Hypothesis #3

Figure 1: In this figure, we illustrate why the MEDIATOR model makes sense for the task of captioned scene understanding. For the caption-image pair (Figure 1a-Figure 1b), we see that parse tree #1 (Figure 1g) shows "standing" (the verb phrase of the noun "dog") connected with "couch" via the "on" preposition, whereas parse trees #2 (Figure 1h) and #3 (Figure 1i) show "woman" connected with "couch" via the "on" preposition. This ambiguity can be resolved if we look at an accurate semantic segmentation such as Hypothesis #3 (Figure 1f) of the associated image (Figure 1b). Likewise, we might be able to do better at semantic segmentation if we choose a segmentation that is consistent with the sentence, such as Segmentation Hypothesis #3 (Figure 1f), which contains a person on a couch with a dog next to them, unlike the other two hypotheses (Figure 1d and Figure 1e).

PASCAL-50S: For PASCAL-50S we first obtained sentences that contain one or more of 7 prepositions (*i.e.*, "with", "next to", "on top of", "in front of", "behind", "by", and "on") that intuitively would typically depend on the relative distance between objects. Then we look for sentences that have preposition phrase attachment ambiguities, *i.e.*, sentences where the parser output has different sets of prepositions for different parsings. Due to our focus on PP attachment, we do not pay attention to other parts of the sentence parse, so the parses can change while the PP attachments remain the same, as in Figure 1h and Figure 1i. The sentences thus obtained are further filtered to obtain sentences in which the objects that are connected by the preposition belonging to one of the 20 PASCAL object categories. Since our vision module is semantic segmentation and not instance-level segmentation, we restrict the dataset to sentences involving prepositions connecting two different PASCAL categories.

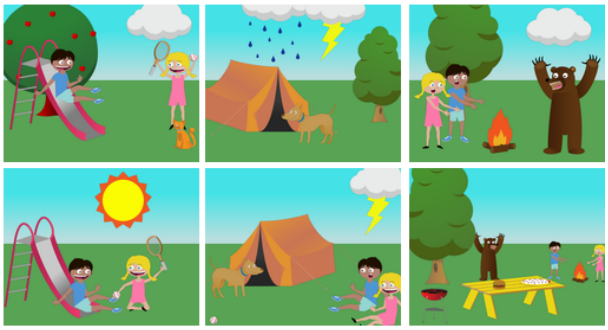


Figure 2: We show some example scenes from (Zitnick and Parikh, 2013). Each column shows two semantically similar scenes, while the different columns show the diversity of scene types.

Thus, our final dataset contains 100 sentences describing 30 unique images and contains 16 of the 20 PASCAL categories as described in the paper. We then manually annotated the ground truth PP attachments. Such manual labeling by student annotators with expertise in NLP takes a lot of time, but results in annotations that are linguistically high-quality, with any inter-human disagreement resolved by strict adherence to rules of grammar.

ABSTRACT-50S: We first obtained sentences that contain one or more of 6 prepositions (*i.e.*, "with", "next to", "on top of", "in front of", "behind", "under"). Due to the semantic differences between the datasets, not all prepositions found in one were present in the other. Further filtering on sentences was done to ensure that the sentences contain at least one preposition phrase attachment ambiguity that is between the clipart noun categories (*i.e.*, each clipart piece has a name, like "snake", that we search the sentence parsing for). This filtering reduced the original dataset of 25,000 sentences and 500 scenes to our final experiment dataset of 399 sentences and 201 scenes. We then manually annotated the ground truth PP attachments.

PASCAL-Context-50S: For PASCAL-Context-50S, we first selected all sentences that have preposition phrase attachment ambiguities. We then plotted the distribution of prepositions in these sentences (see Figure 3). We found that there was a drop in the percentage of sentences for prepositions that appear in the sorted list after "down". Therefore, we only kept sentences that have one or more 2-D visual prepositions in the list of prepositions up to "down". Thus we ended up with the following 7 prepositions: "on", "with", "next to", "in front of", "by", "near", and "down". We then further sampled sentences to ensure uniform distribution across prepositions. Unlike PASCAL-50S, we did not filter sentences based on whether the objects connected by the prepositions belong to one of 60 PASCAL Context categories or not. Instead, we used the word2vec (Mikolov et al., 2013) similarity between the objects in the sentence and the PASCAL Context categories as one of the features. Thus, our final dataset contains 1,822 sentences describing 966 unique images.

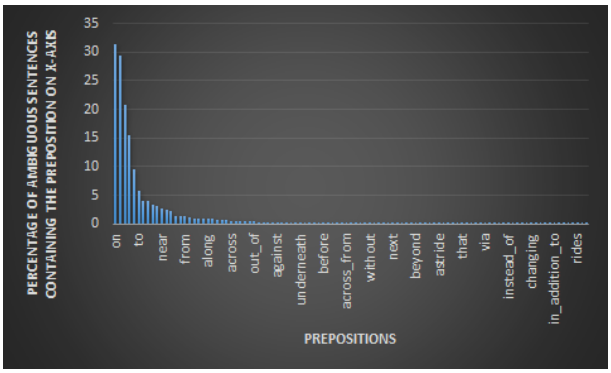


Figure 3: We show the percentage of ambiguous sentences in PASCAL-Context-50S dataset before filtering for prepositions. We found that there was a drop in the percentage of sentences for prepositions that appear in the sorted list after "down". So, for the PASCAL-Context-50S dataset we only keep sentences that have one or more visual prepositions in the list of prepositions up to "down".

The ground truth PP attachments for these 1,822 sentences were annotated by AMT workers. For each unique prepositional relation in a sentence, we showed the workers the prepositional relation of the form primary object preposition secondary object and its associated image and sentence and asked them to specify whether the prepositional relation is correct or not correct. We also asked them to choose the third option - "Primary object/ secondary object is not a noun in the caption" in case that happened. The user interface used to collect these annotations is shown in Figure 4. We collected five answers for each prepositional relation. For evaluation, we used the majority response. We found that 87.11% of human responses agree with the majority response, indicating that even though AMT workers were not explicitly trained in rules of grammar by us, there is relatively high inter-human agreement.

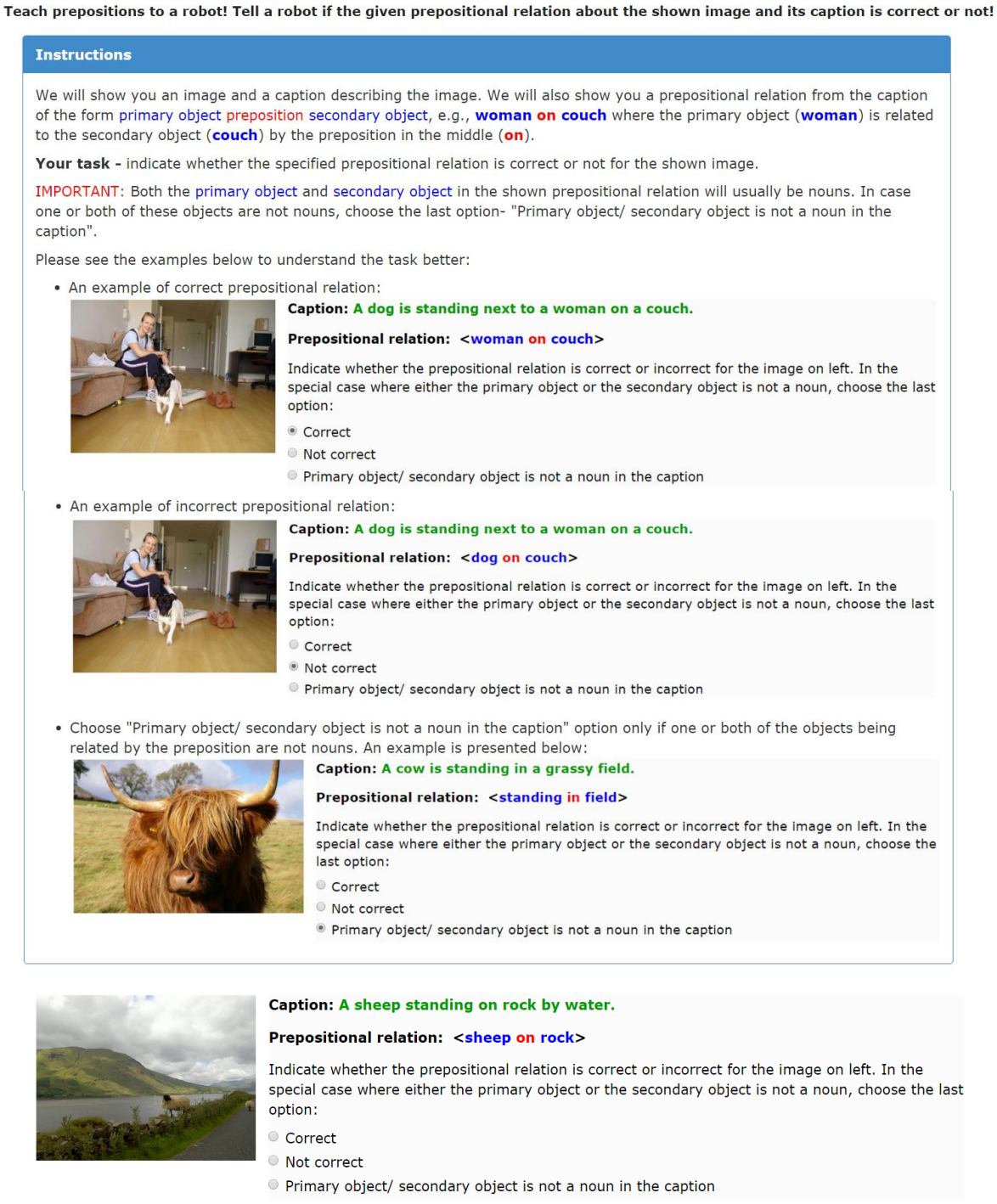
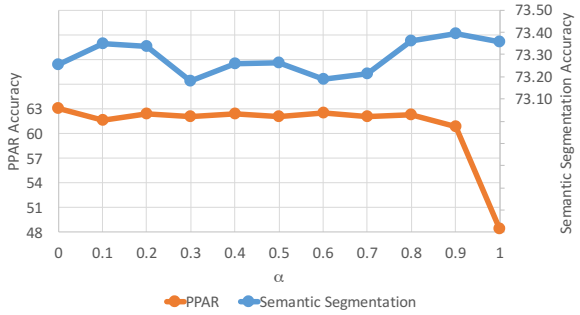


Figure 4: The AMT interface to collect ground truth annotations for prepositional relations. Five answers were collected for each prepositional relation. The majority response is used for evaluation. The AMT workers are asked to select if the preposition is correct, not correct, or that the primary or secondary object is not a noun in the caption. Examples for all three answer choices are shown in the instructions presented to the workers.

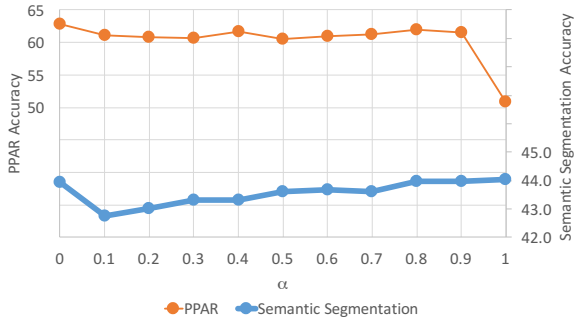
	"on"	"with"	"next to"	"in front of"	"by"	"near"	"down"
Acc.	64.26	63.30	60.98	56.86	62.81	67.83	67.23
Gain	12.47	15.41	11.18	13.27	14.13	12.82	17.16

Table 1: Performances and gains over the independent baseline on PASCAL-Context-50S for each preposition.

4 Effect of Different Weighting of Modules



(a) PASCAL-50S



(b) PASCAL-Context-50S

Figure 6: Accuracies MEDIATOR (both modules) vs α, where α is the coefficient for the semantic segmentation module and 1-α is the coefficient for the PPAR resolution module in the loss function. Our approach is fairly robust to the setting of α, as long as it is not set to either extremes, since that limits the synergy between the modules. (a) shows the results for PASCAL-50S, and (b) shows the results for PASCAL-Context-50S.

So far we have used the "natural" setting of $\alpha = 0.5$, which gives equal weight to both modules. Note that α is not a parameter of our approach; it is a design choice that the user/experiment-designer makes. To see the effect of weighting the modules differently, we tested our approach for various values of α. Figure 6 shows how the accuracies of each module vary depending on α for the MEDIATOR model for PASCAL-50S and PASCAL-Context-50S. Recall that α is the coefficient for the semantic segmentation module and 1-α is the coefficient for the PPAR resolution module in the loss function. We see that as expected, putting no or little weight on the PPAR module drastically hurts performance for that module. Our approach is fairly robust to the setting of α, with a peak lying but any weight on it performs fairly similar with the peak lying somewhere between the extremes. The segmentation module has similar behavior, though it is not as sensitive to the choice of α. We believe this is because of small "dynamic range" of this module – the gap between the 1-best and oracle segmentation is smaller and thus the MEDIATOR can always default to the 1-best as a safe choice.

5 Performances for Each Preposition

We provide performances and gains over the independent baseline on PASCAL-Context-50S for each preposition in Table 1. We see that vision helps all prepositions.

6 Qualitative Examples

Figure 7 - Figure 12 show qualitative examples for our experiments. Figure 7 - Figure 9 show examples for the multiple modules examples (semantic segmentation and PPAR), and Figure 10 - Figure 12 show examples for the single module experiment. In each figure, the top row shows the image and the associated sentence. For the multiple modules figures, the second and third row show the diverse segmentations of the image, and the bottom two rows show different parsings of the sentence (last two rows for single module examples, as well). In these examples our approach uses 10 diverse solutions for the semantic segmentation module and 10 different solutions for the PPAR module. The highlighted pairs of solutions show the solutions picked by the MEDIATOR model. Examining the results can give you a sense of how the parsings can help the semantic segmentation module pick the best solution and vice-versa.

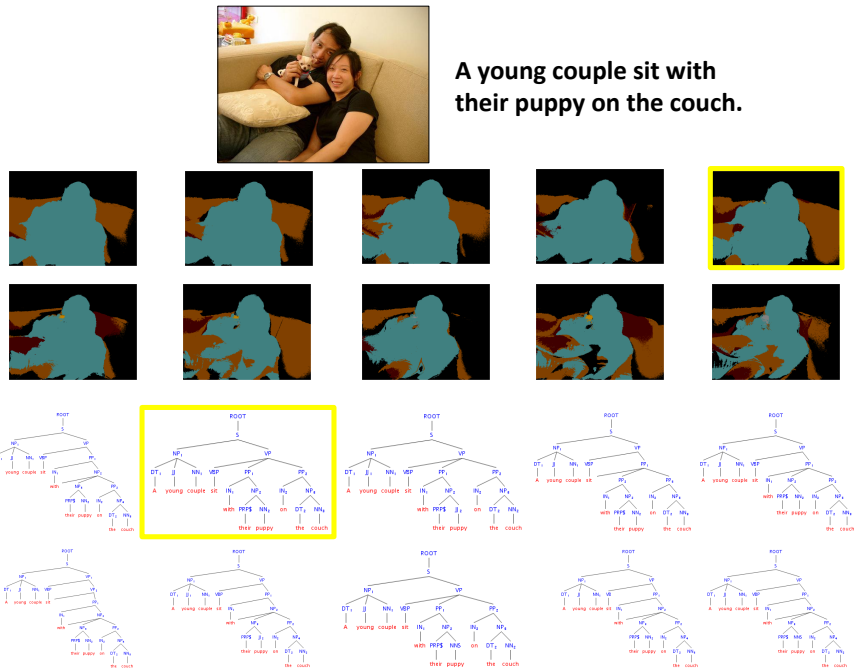


Figure 7: Example 1 – multiple modules (SS and PPAR).

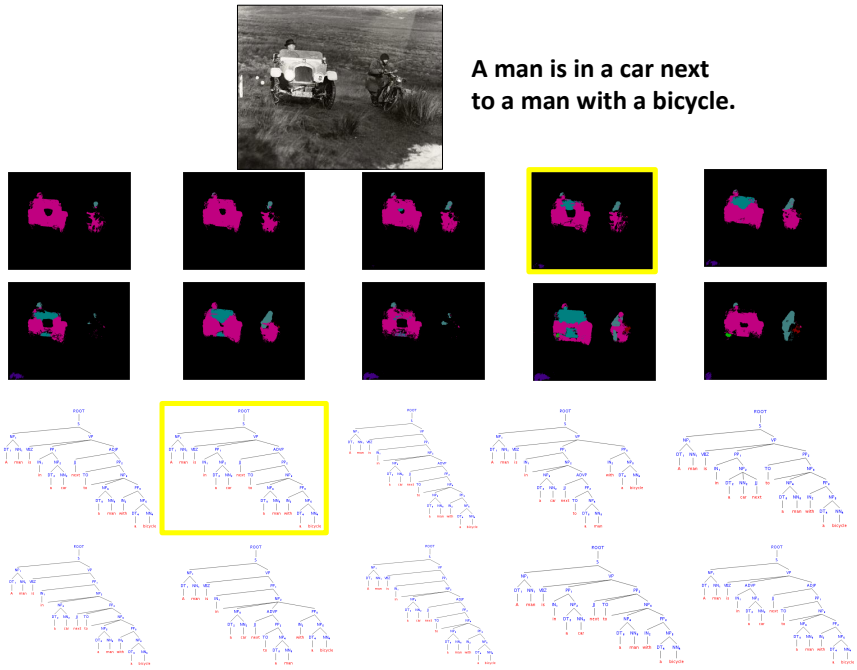


Figure 8: Example 2 – multiple modules (SS and PPAR).

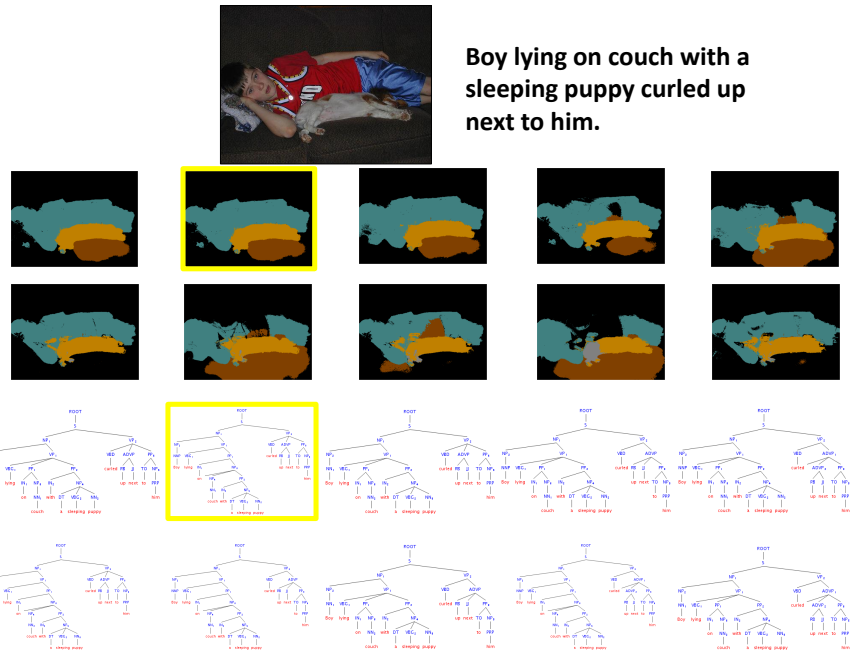


Figure 9: Example 3 – multiple modules (SS and PPAR).

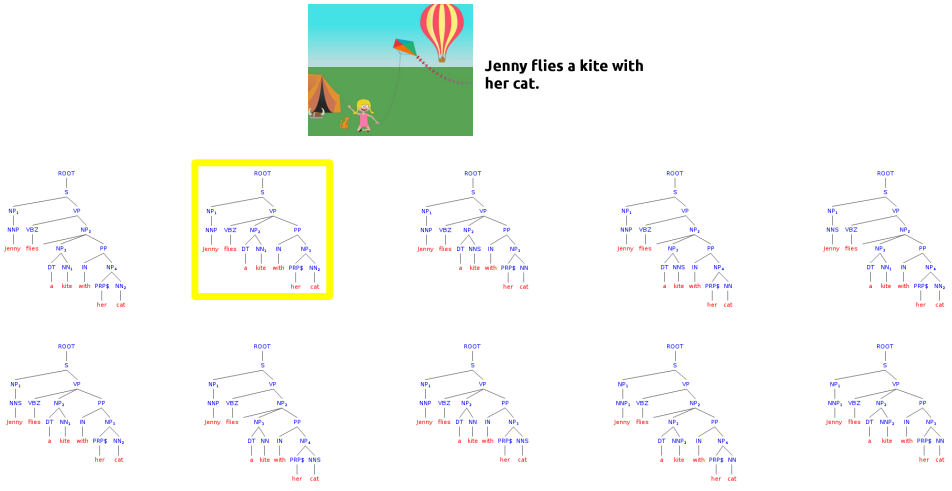


Figure 10: Example 1 – single module (PPAR).

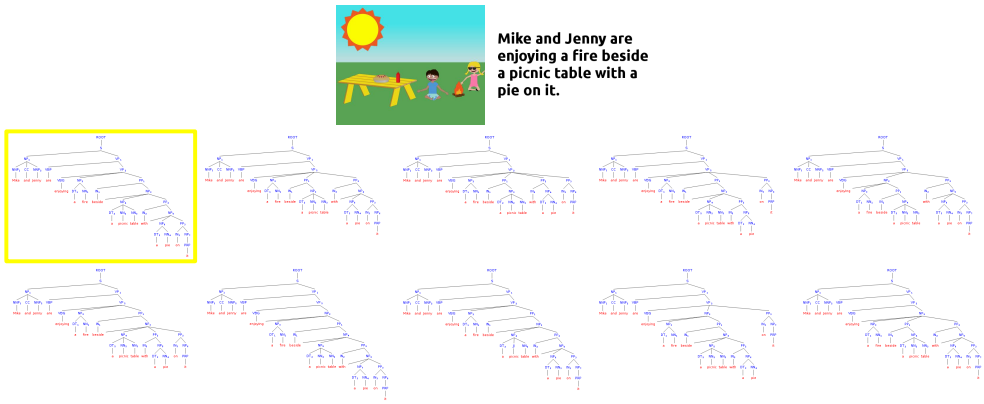


Figure 11: Example 2 – single module (PPAR).

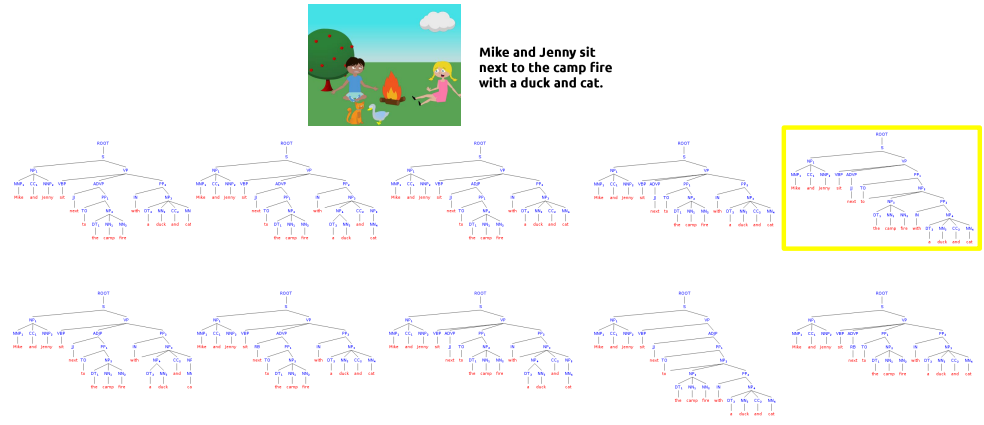


Figure 12: Example 3 – single module (PPAR).

References

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *ICLR*.

C. Lawrence Zitnick and Devi Parikh. 2013. Bringing Semantics Into Focus Using Visual Abstraction. In *CVPR*.