# Supplementary Material for "Learning to Assemble Neural Module Tree Networks for Visual Grounding"

Daqing Liu[1] Hanwang Zhang[2] Feng Wu[1] Zheng-Jun Zha[1]
[1]University of Science and Technology of China, [2]Nanyang Technological University
liudq@mail.ustc.edu.cn, hanwangzhang@ntu.edu.sg, fengwu@ustc.edu.cn, zhazj@ustc.edu.cn

## 1. Implementation of Tree LSTM

We simplified the implementation of tree LSTM (Eq. (4)) in the main paper as:

$$\boldsymbol{c}_t^{\uparrow}, \boldsymbol{h}_t^{\uparrow} = \text{TreeLSTM}(\boldsymbol{e}_t, \{\boldsymbol{c}_{tj}^{\uparrow}\}, \{\boldsymbol{h}_{tj}^{\uparrow}\}), \quad j \in \mathcal{C}_t, \tag{1}$$

where $\boldsymbol{c}_{tj}^{\uparrow}$, $\boldsymbol{h}_{tj}^{\uparrow}$ denote the cell and hidden vectors of the $j$-th child of node $t$. Specifically, our tree LSTM transition equations are:

$$\tilde{\boldsymbol{h}}_t = \sum\nolimits_{j \in \mathcal{C}_t} \boldsymbol{h}_{tj}^{\uparrow}, \tag{2}$$

$$\boldsymbol{i}_t = \sigma(W^{(i)}\boldsymbol{e}_t + U^{(i)}\tilde{\boldsymbol{h}}_t + \boldsymbol{b}^{(i)}), \tag{3}$$

$$\boldsymbol{o}_t = \sigma(W^{(o)}\boldsymbol{e}_t + U^{(o)}\tilde{\boldsymbol{h}}_t + \boldsymbol{b}^{(o)}), \tag{4}$$

$$\boldsymbol{u}_t = \tanh(W^{(u)}\boldsymbol{e}_t + U^{(u)}\tilde{\boldsymbol{h}}_t + \boldsymbol{b}^{(u)}), \tag{5}$$

$$\boldsymbol{f}_{tj} = \sigma(W^{(f)}\boldsymbol{e}_t + U^{(f)}\boldsymbol{h}_{tj}^{\uparrow} + \boldsymbol{b}^{(f)}), \tag{6}$$

$$\boldsymbol{c}_t^{\uparrow} = \boldsymbol{i}_t \odot \boldsymbol{u}_t + \sum\nolimits_{j \in \mathcal{C}_t} \boldsymbol{f}_{tj} \odot \boldsymbol{c}_{tj}^{\uparrow}, \tag{7}$$

$$\boldsymbol{h}_t^{\uparrow} = \boldsymbol{o}_t \odot \tanh(\boldsymbol{c}_t^{\uparrow}), \tag{8}$$

where $\odot$ is the element-wise multiplication, $\sigma(\cdot)$ is the sigmoid function, $W$, $U$, $b$ are trainable parameters.

## 2. More Qualitative Results

In this section, we provide more qualitative results to demonstrate the internal reasoning steps of NMTREE. In Figure 1, we visualize the reasoning process inside `Comp` modules. In Figure 2, Figure 3, and Figure 4, we visualize the tree structures, the module assembly, the attention map at each intermediate step, and the final results. Specifically, Figure 2 are qualitative results with ground-truth bounding boxes. As comparison, we also show some failure cases. Figure 3 are qualitative results with detected bounding boxes. Figure 4 are qualitative results with detected masks.

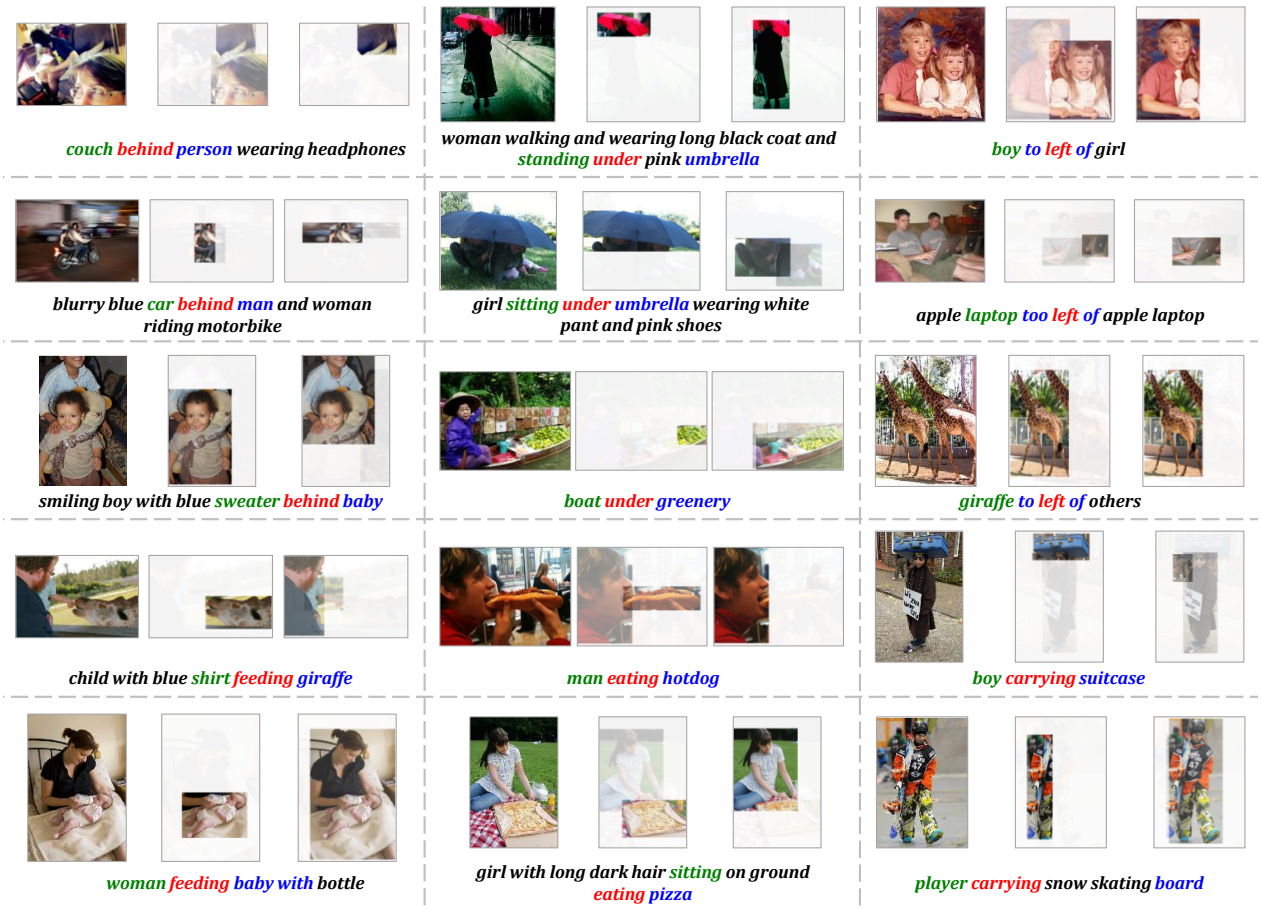
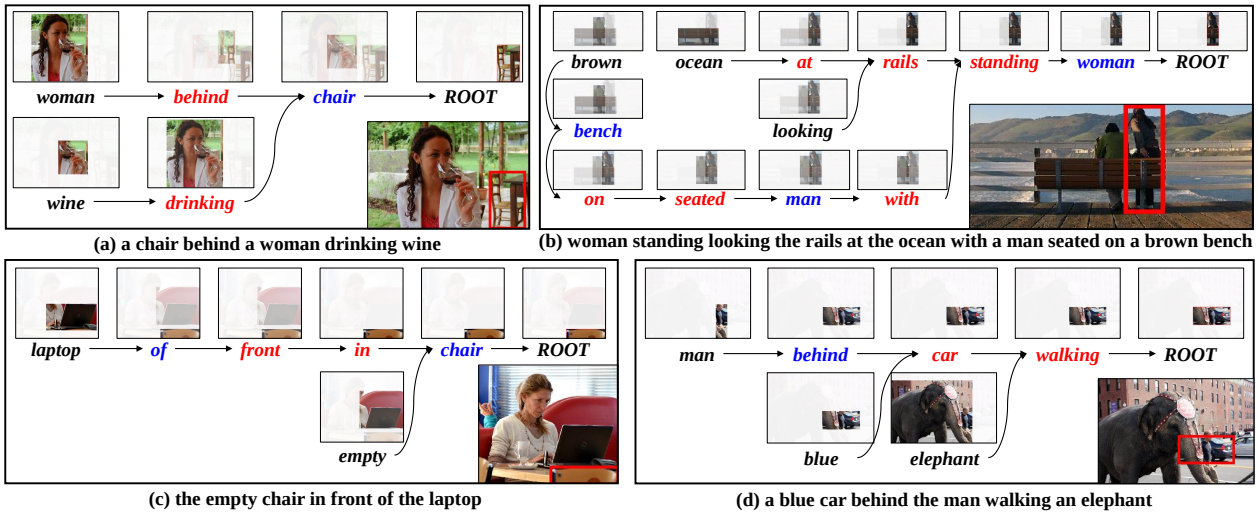



Figure 1. The compositional reasoning inside Comp. Each example contains the original image (left), the contextual attention map (middle) and the output attention map (right). We represent partial tree structure by colors: red for the current node, blue for children and green for parent.



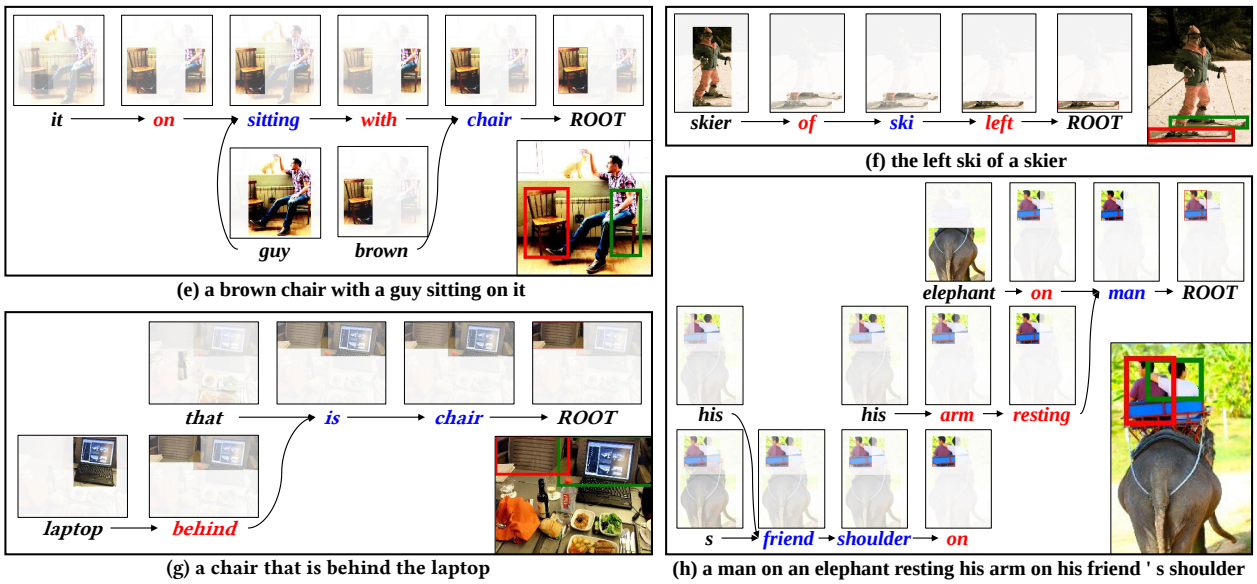

(a) a chair behind a woman drinking wine

(b) woman standing looking the rails at the ocean with a man seated on a brown bench

(c) the empty chair in front of the laptop

(d) a blue car behind the man walking an elephant

(a) correct




(e) a brown chair with a guy sitting on it

(f) the left ski of a skier

(g) a chair that is behind the laptop

(h) a man on an elephant resting his arm on his friend ' s shoulder

(b) incorrect

Figure 2. Qualitative results with ground-truth bounding boxes. Words in different colors indicate corresponding modules: black for `Single`, red for `Comp`, and blue for `Sum`. The bottom right corner is the original image with a green bounding box as ground-truth and a red bounding box as our result.

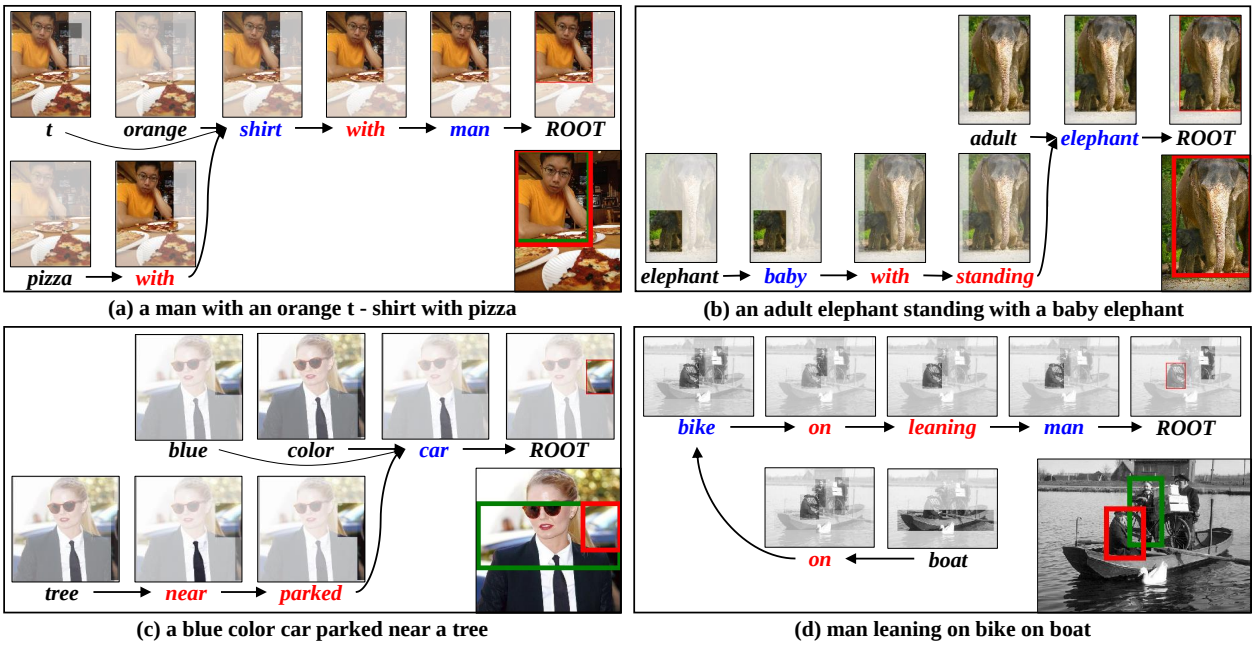



Figure 3. Qualitative results with detected bounding boxes. Words in different colors indicate corresponding modules: black for `Single`, red for `Comp`, and blue for `Sum`. The bottom right corner is the original image with a green bounding box as ground-truth and a red bounding box as our result.

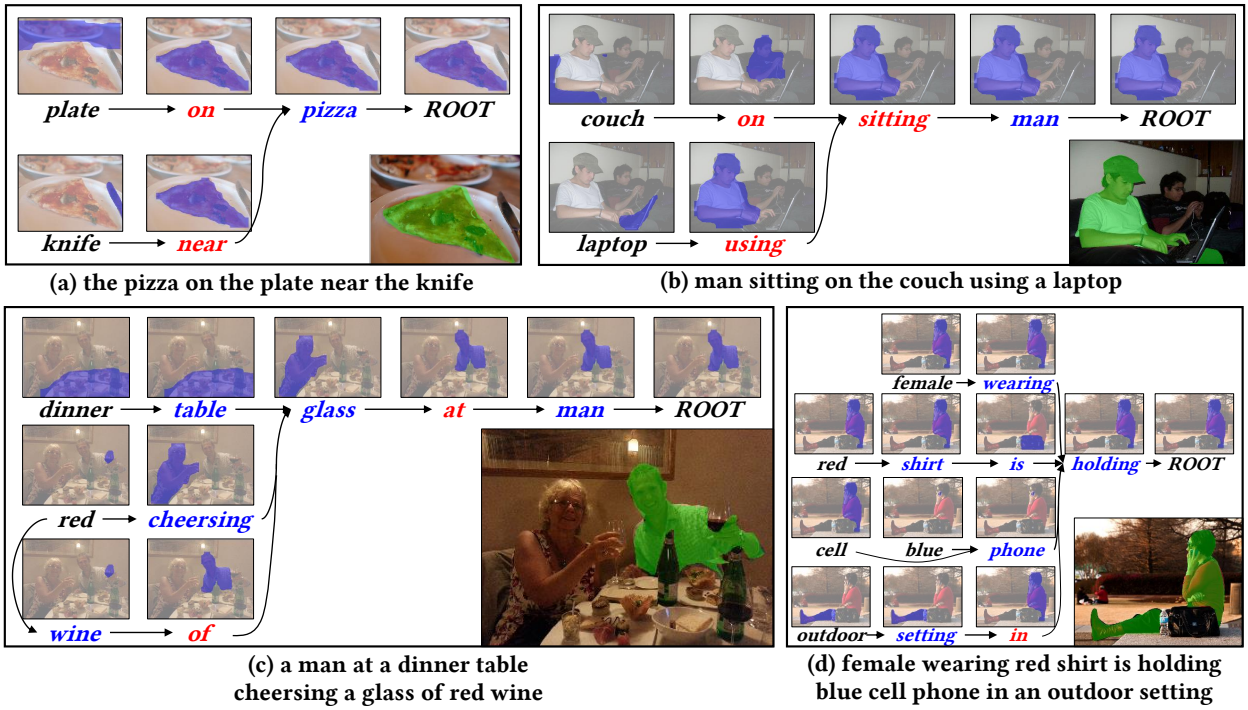



Figure 4. Qualitative results with detected masks. Words in different colors indicate corresponding modules: black for `Single`, red for `Comp`, and blue for `Sum`. The blue masks indicate the regions with maximum score, and the green masks indicate the ground-truth.