Comprehensive Image Captioning via Scene Graph Decomposition

Yiwu Zhong[1], Liwei Wang[2], Jianshu Chen[2], Dong Yu[2], Yin Li[1]

[1] University of Wisconsin-Madison, United States

[2] Tencent AI Lab, Bellevue, United States

Image Captioning

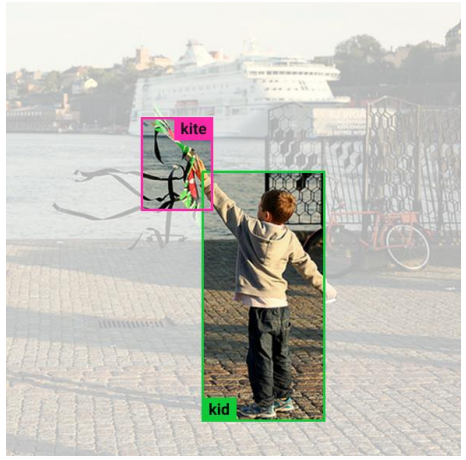
Deep Model

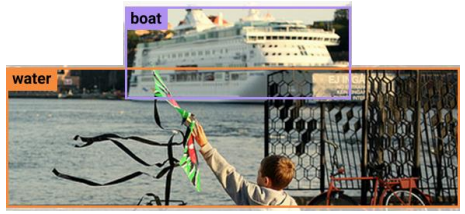
A young boy is flying a kite. **Accurate Captioning**

A young boy is flying a kite.

…

A kite is flying over the boy. **Diverse Captioning**

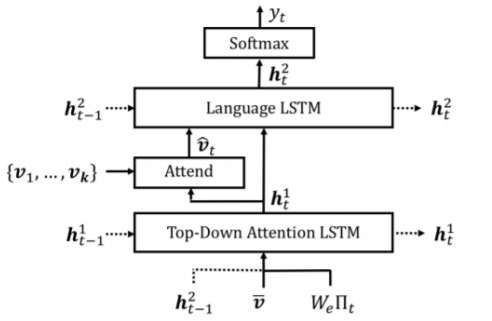




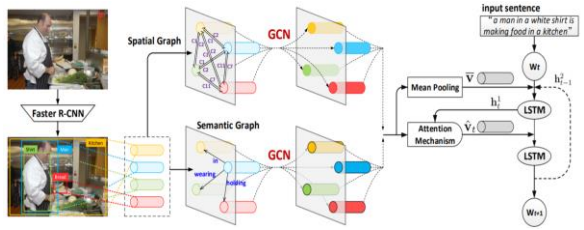
A young boy is flying a kite. **Grounded Captioning**

A ship is sailing on the river. **Controllable Captioning**

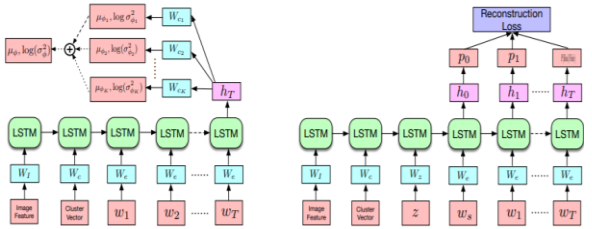
Related Work



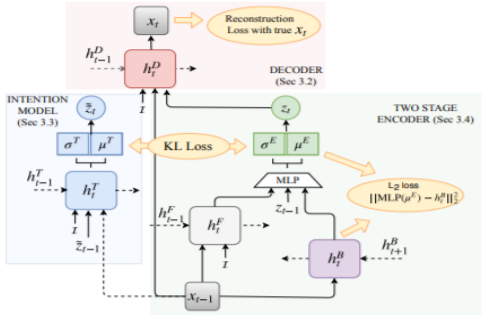
Anderson et al., CVPR 2018



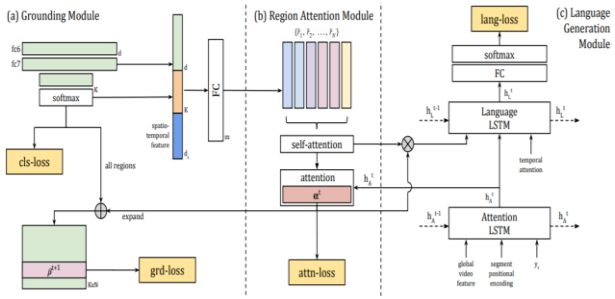
Yao et al., ECCV 2018



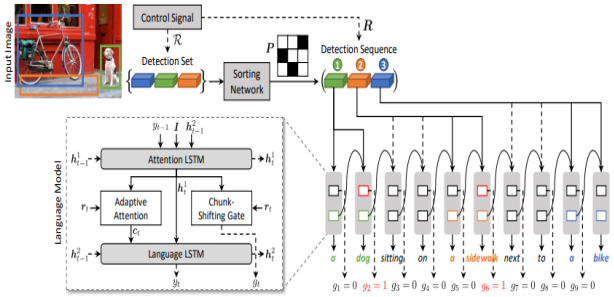
Wang et al., NeurIPS 2017



Aneja et al., ICCV 2019



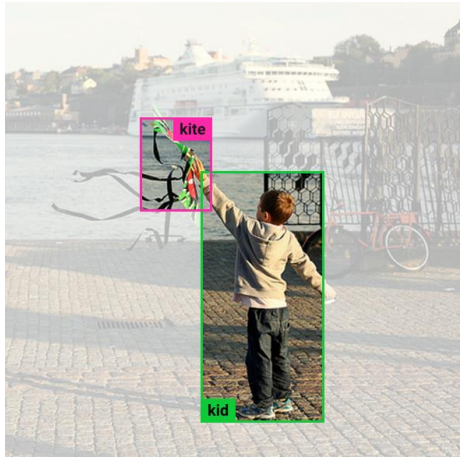
Zhou et al., CVPR 2019



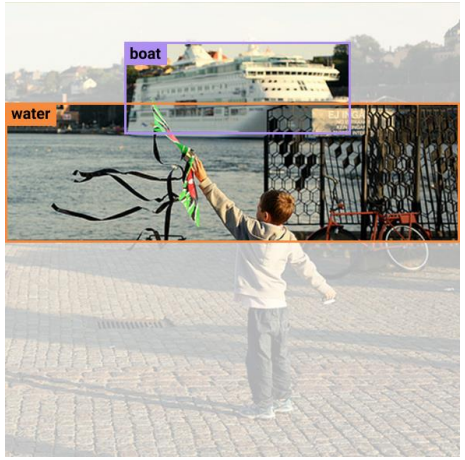
Cornia et al., CVPR 2019

Comprehensive Image Captioning

Our Model



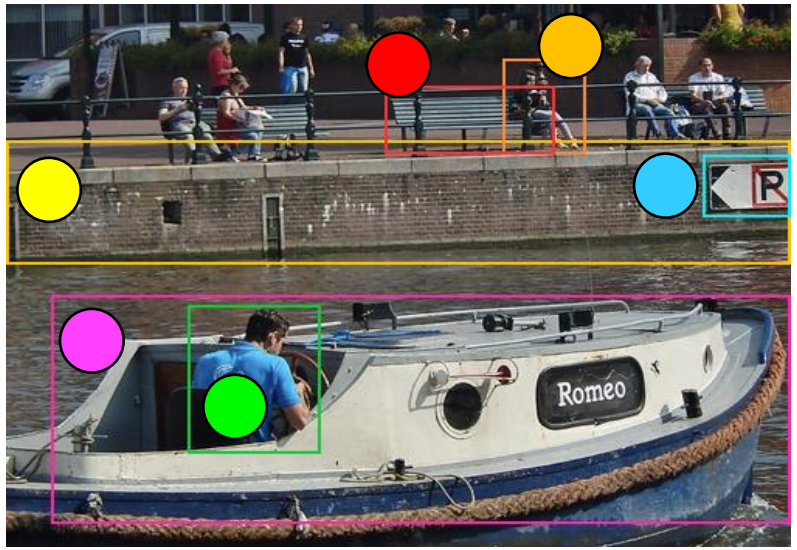
A young boy is flying a kite.

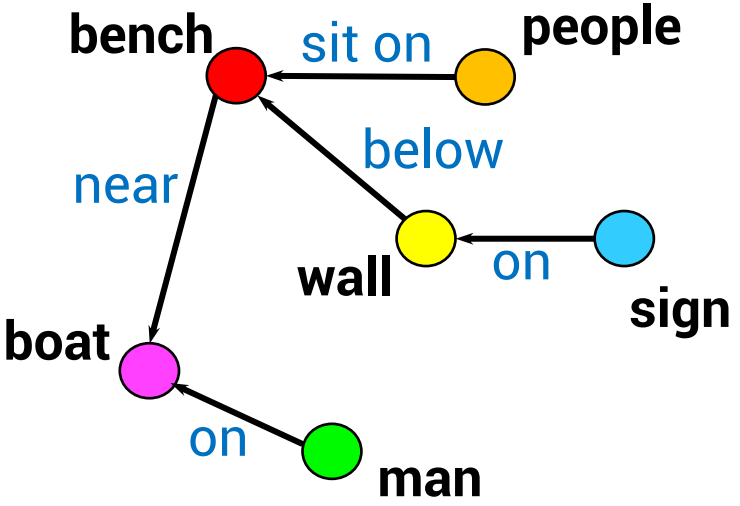


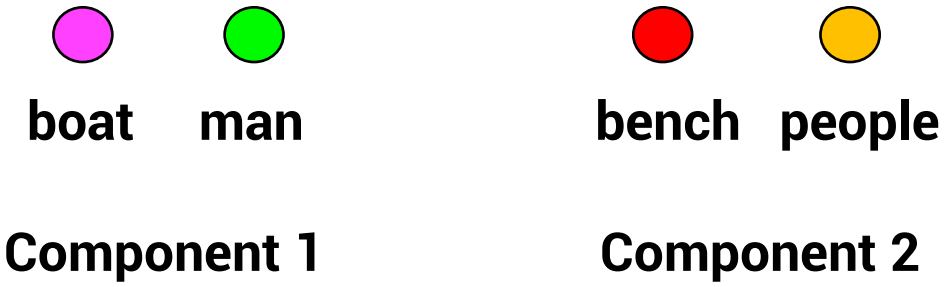
A cruise ship is sailing on the river.

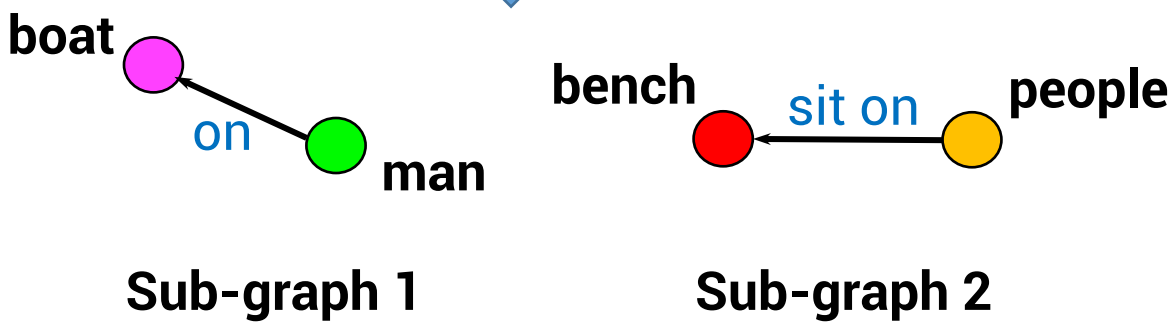
A bike is parked on the street.

Image Components & Image Scene Graph





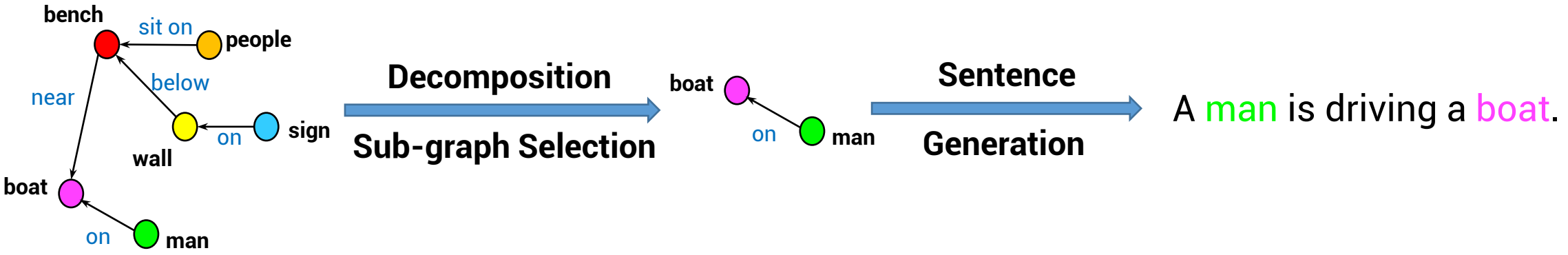




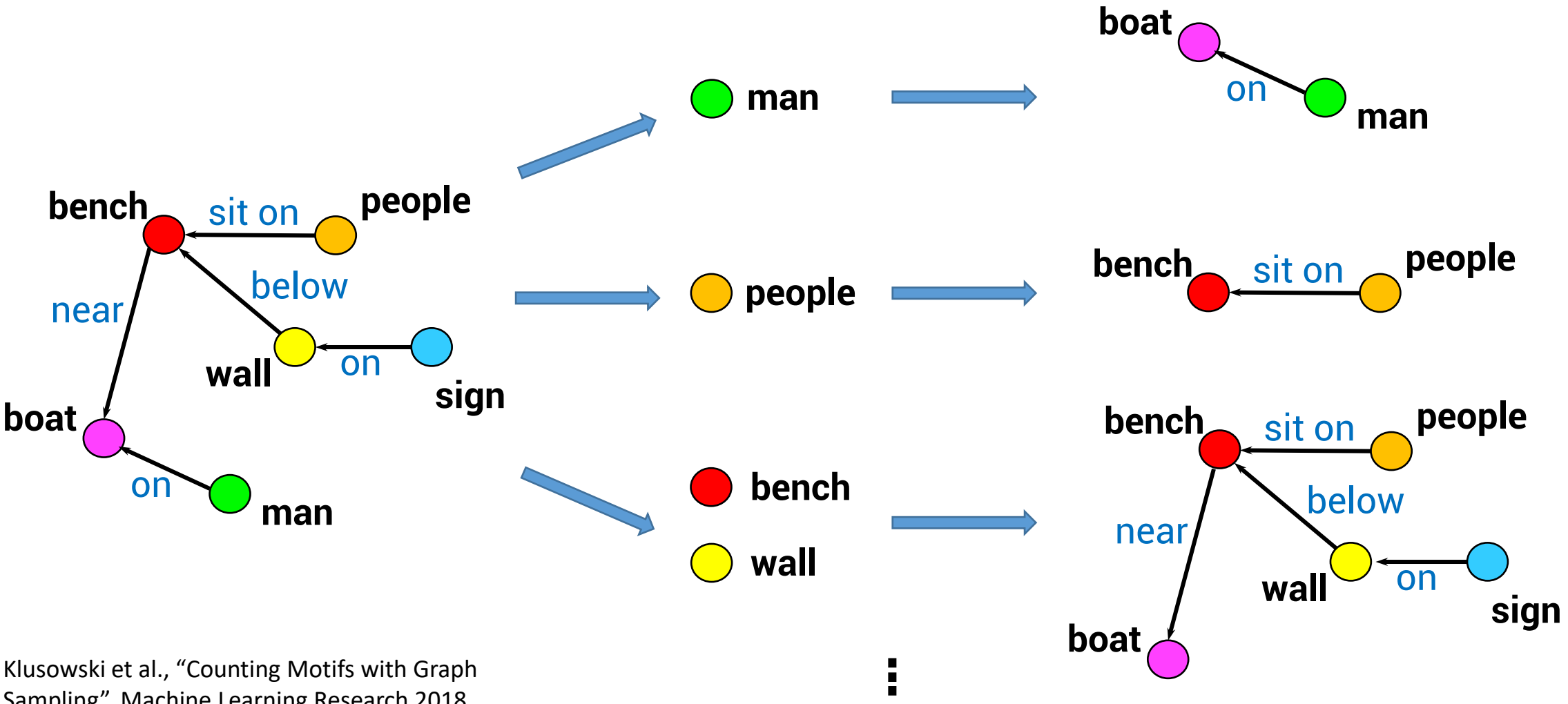
Zellers et al., “Neural Motifs: Scene Graph Parsing with Global Context”, CVPR 2018

Key Idea: Captions from Sub-Graphs

- Decomposing scene graph into sub-graphs
- Selecting a meaningful sub-graph to decode a sentence



Scene Graph Decomposition

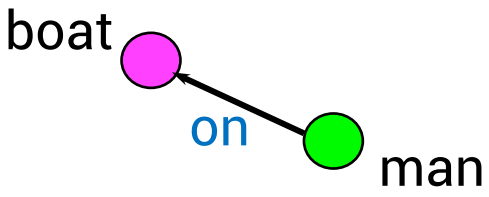


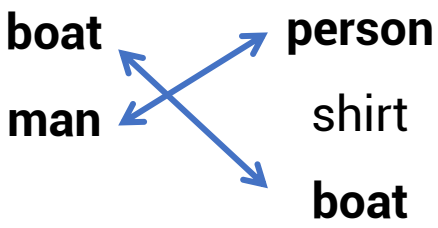
Klusowski et al., “Counting Motifs with Graph Sampling”, Machine Learning Research 2018

Identifying Meaningful Sub-graphs

Meaningful Sub-graphs:

The sub-graphs that can be matched to the ground truth captions.





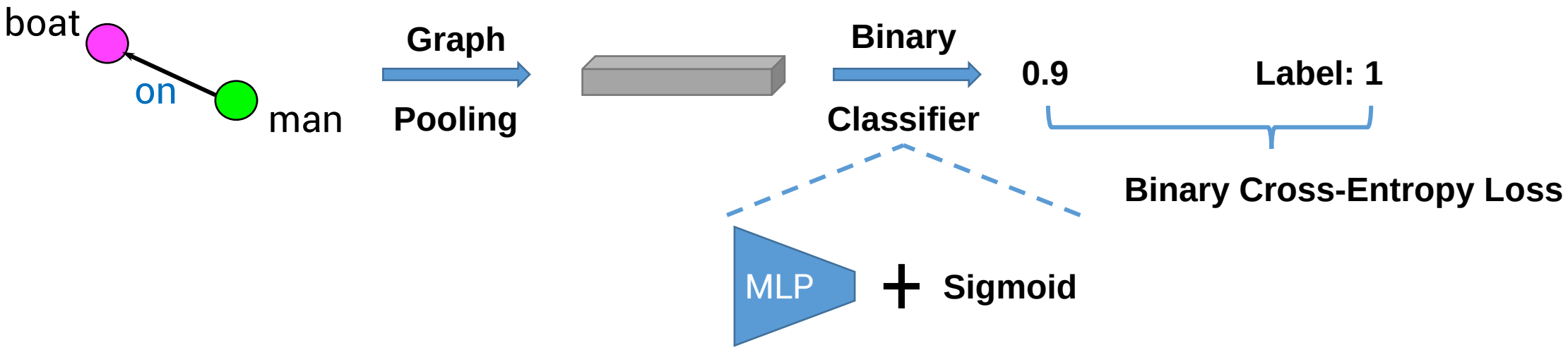
"A person in a blue shirt is driving a boat."

$$Intersection\ of\ Union \xrightarrow{> \text{threshold}} \textbf{Label = 1}$$

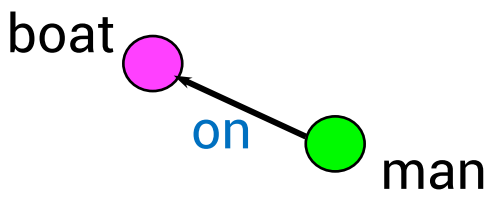
Sub-graph Proposal Network (sGPN)

Goal:

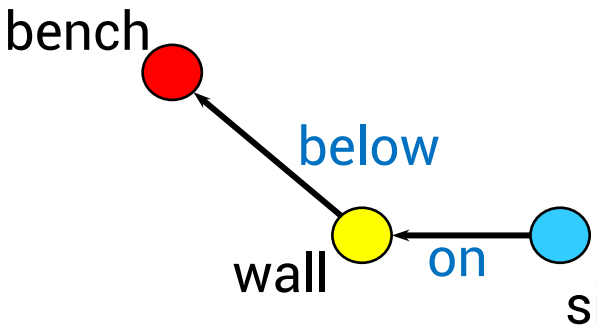
Design a binary classifier to identify the meaningful sub-graphs.



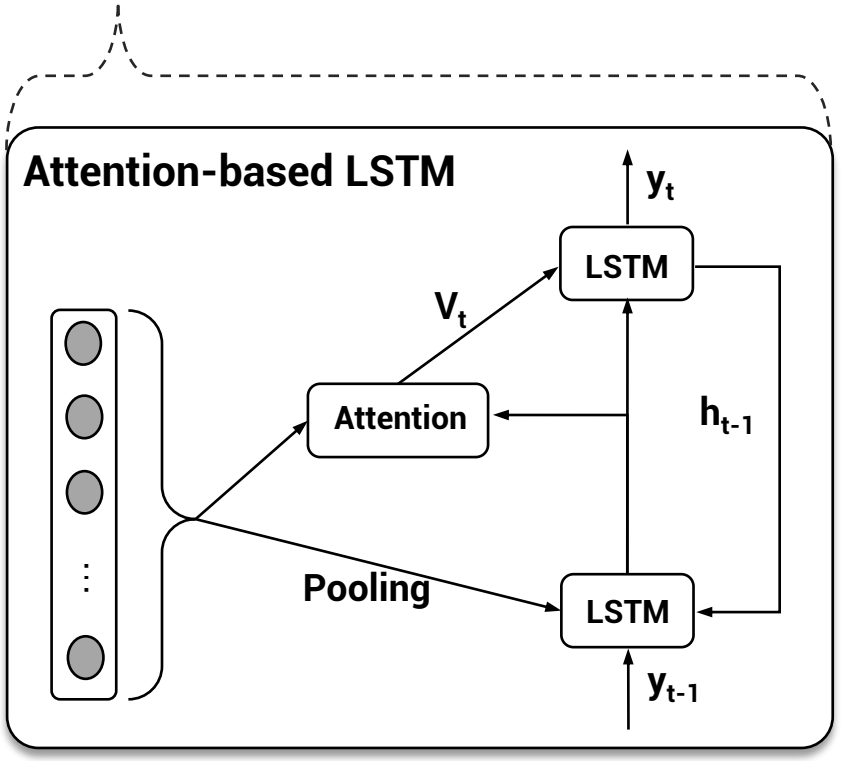
Sub-graph Decoding



sGPN → 0.9 → LSTM → A man is driving a boat.

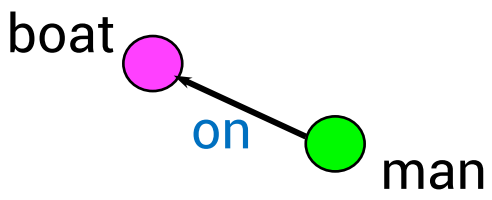


sGPN → 0.2

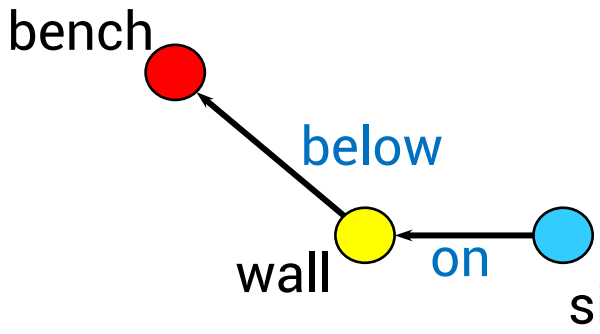


Anderson et al., "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering", CVPR 2018

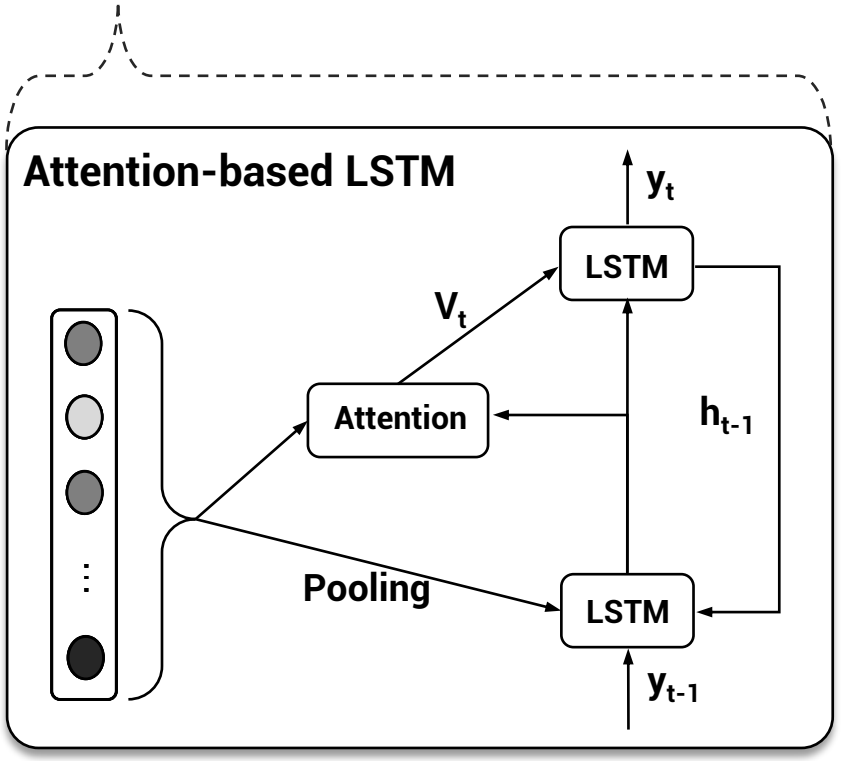
Grounded Caption



sGPN → 0.9 → LSTM → A man is driving a boat.

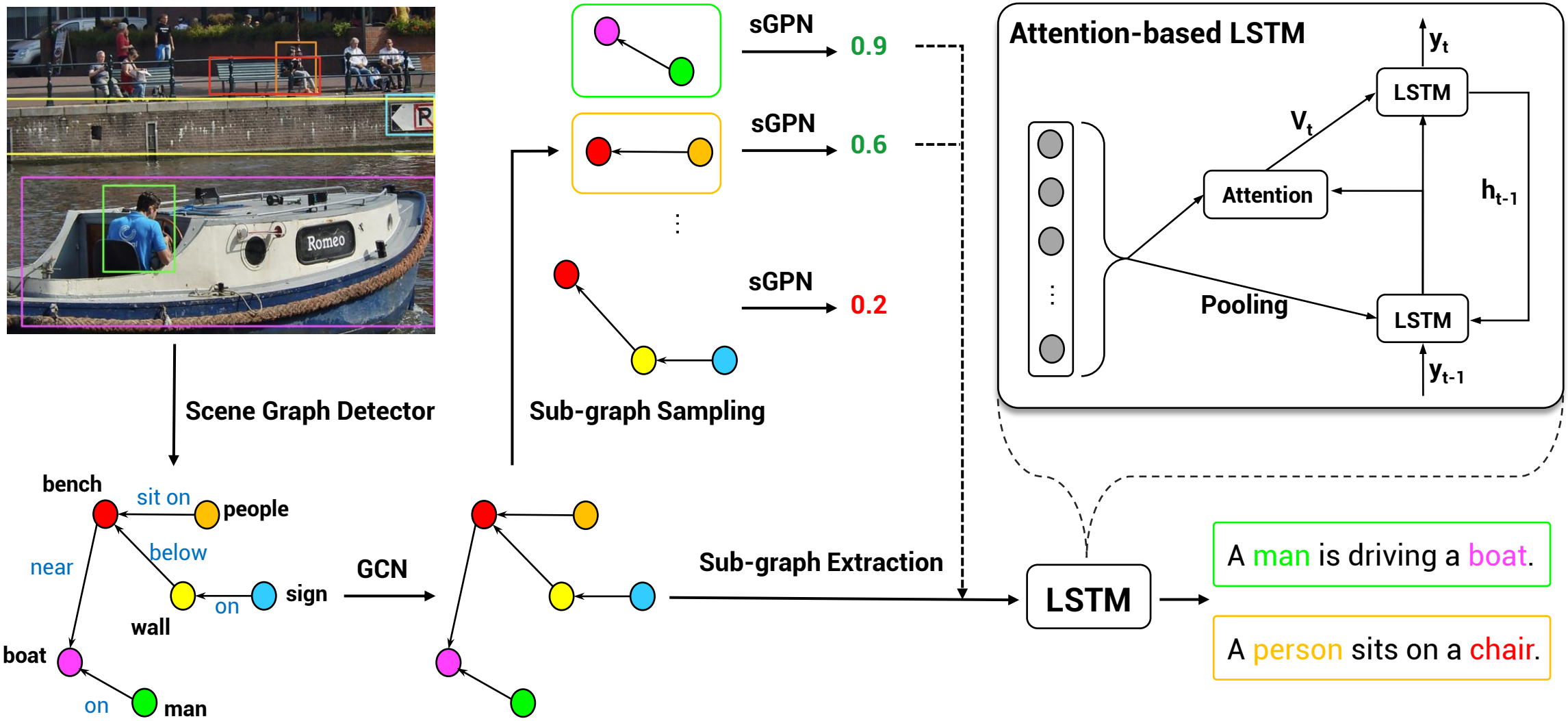


sGPN → 0.2

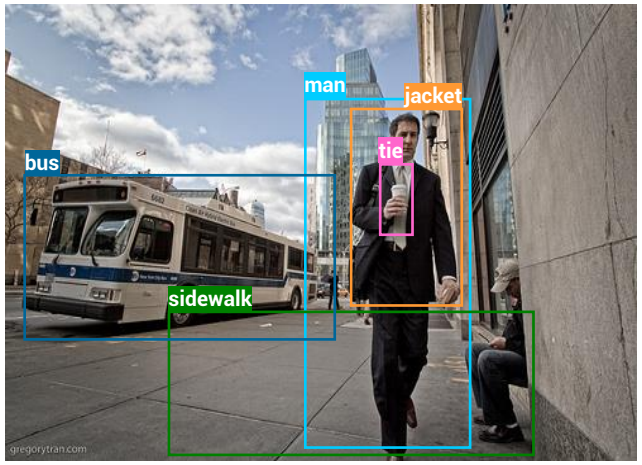


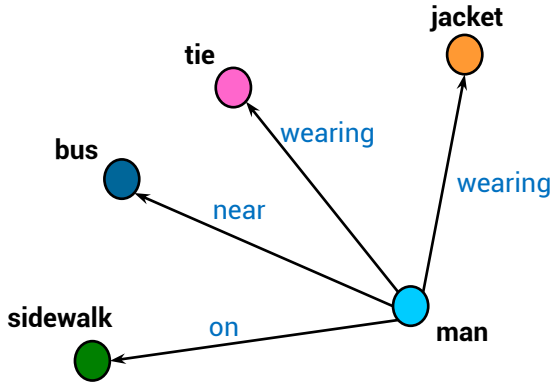
Anderson et al., "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering", CVPR 2018

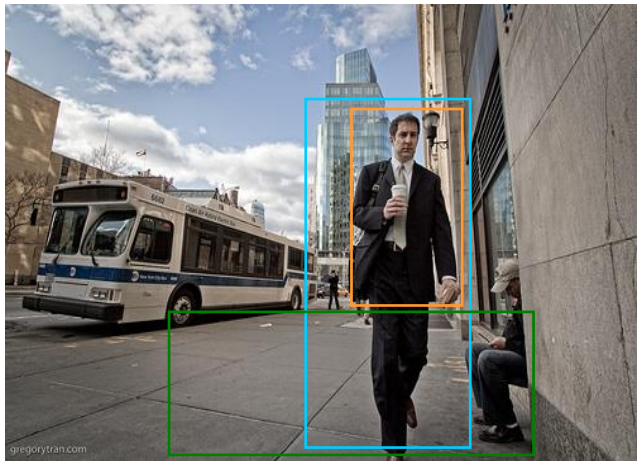
Sub-graph Captioning



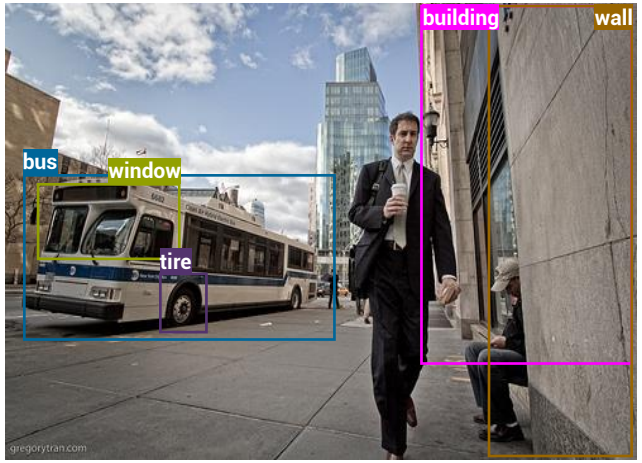
Qualitative Results

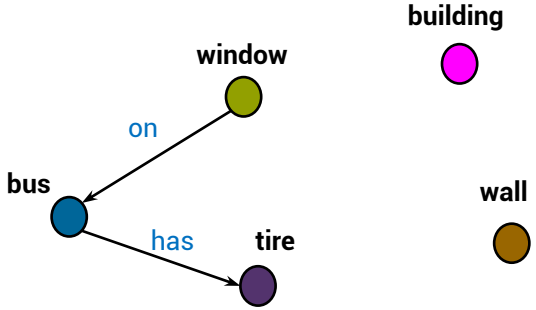


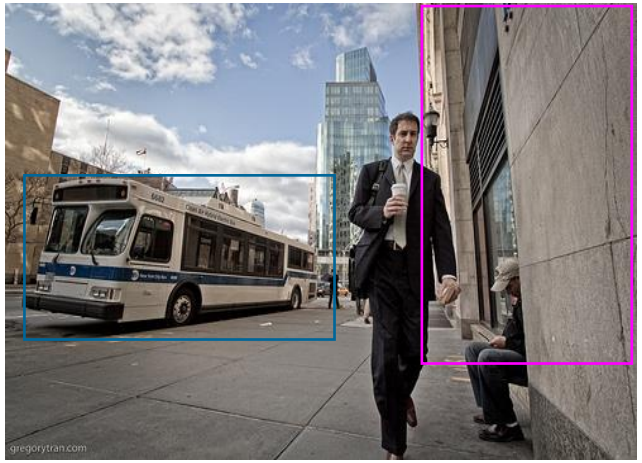




A man in a suit is walking down the street.

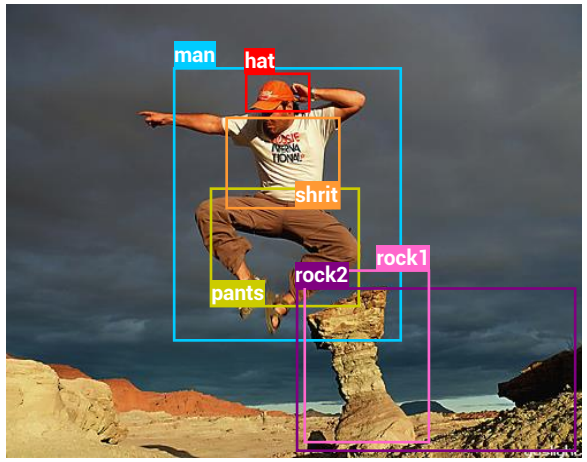


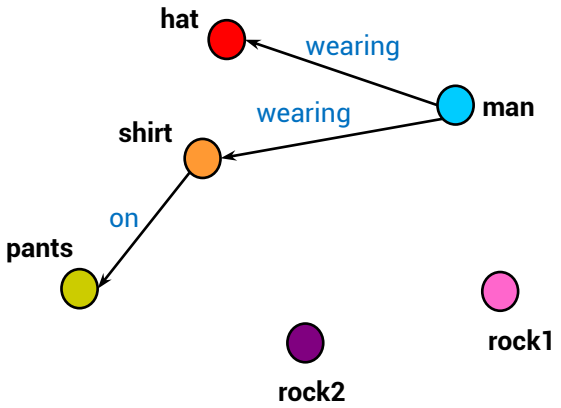




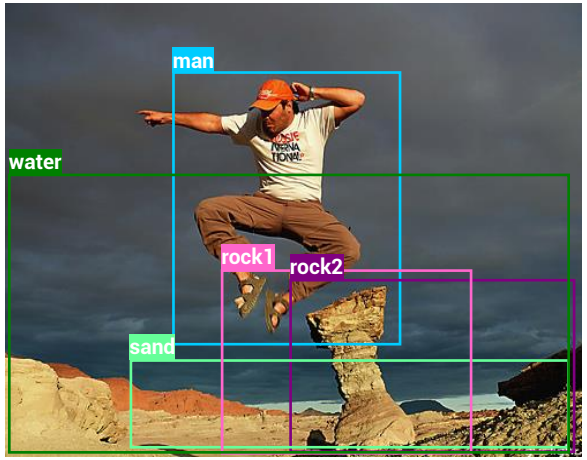
A bus is parked in front of a building.

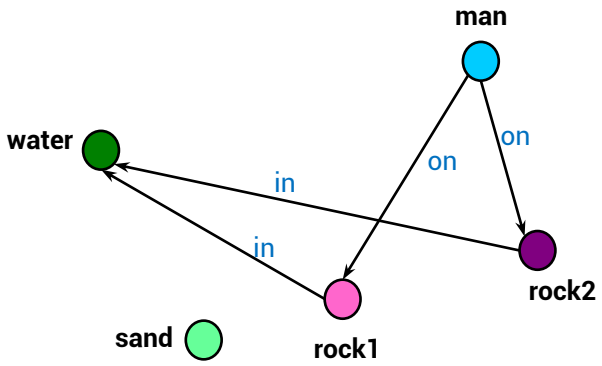
Qualitative Results





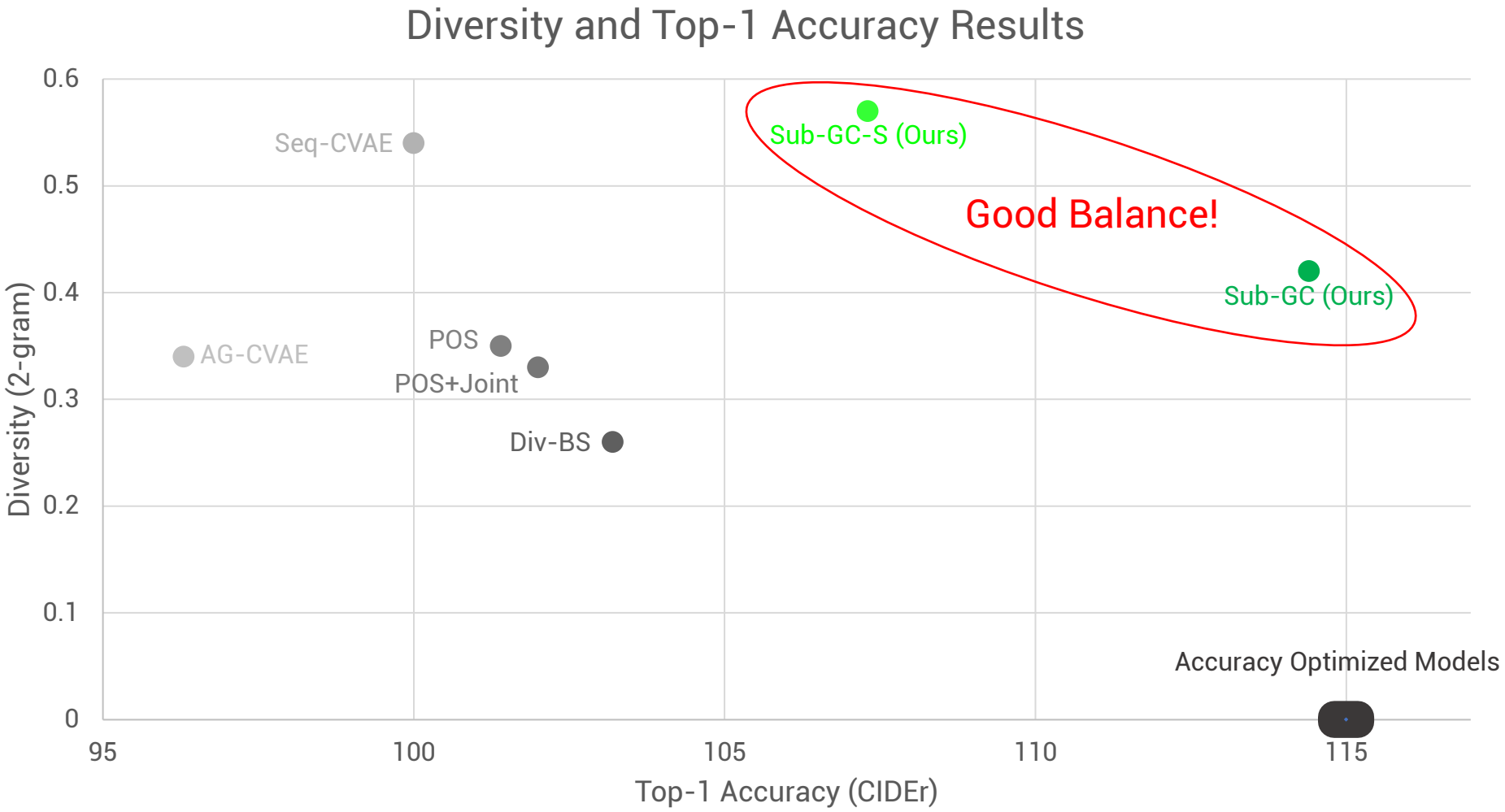
A man in a orange hat and brown pants is jumping off a rock.



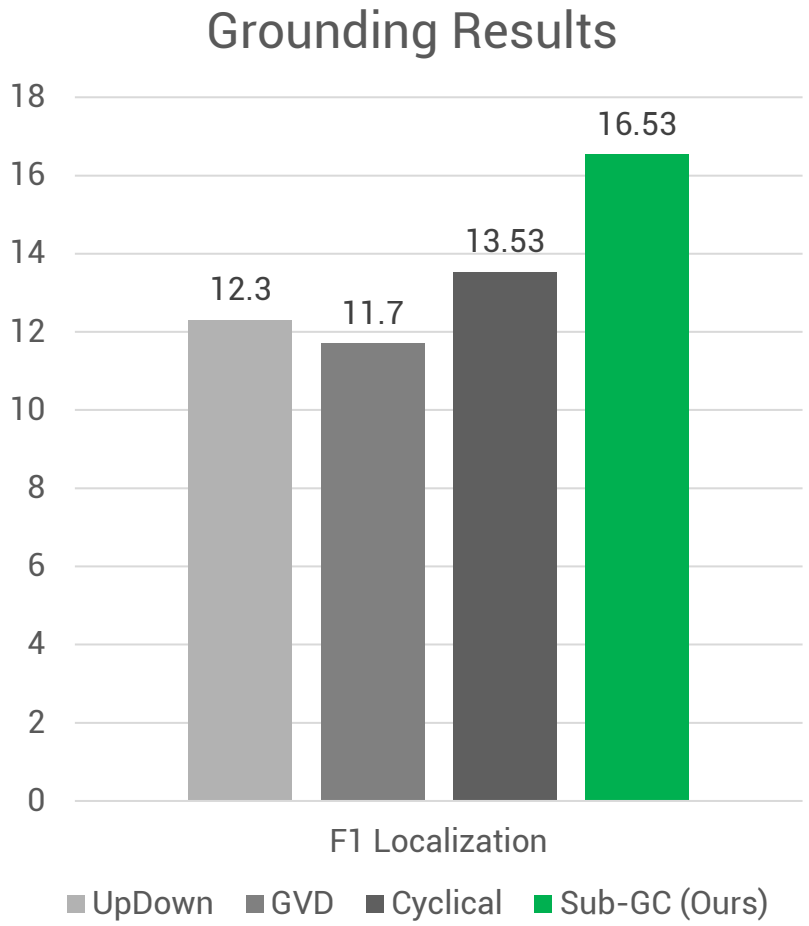


A man is jumping off a rock in a rocky area.

Results - Diverse and Accurate Captioning



Results - Grounded Captioning

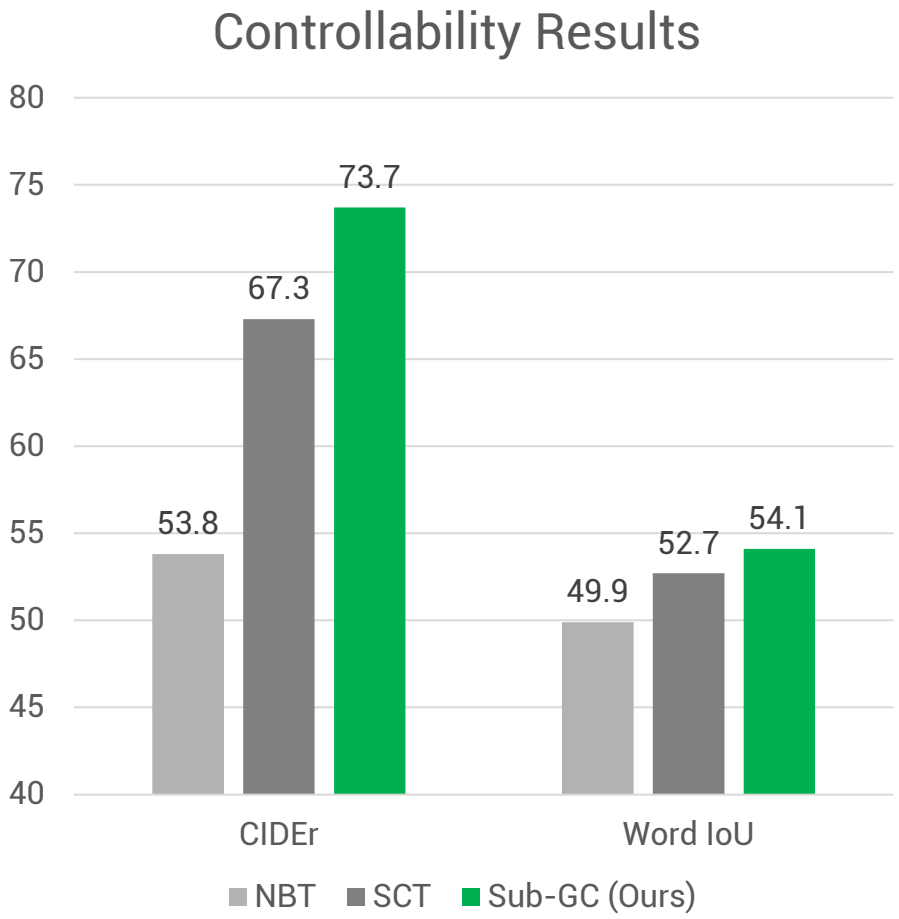


Grounding: locate image regions associated with sentence tokens

Metric: F1 score for localization

Zhou et al., “Grounded Video Description”, CVPR 2019

Results - Controllable Captioning



Controllability: decode a target sentence given a set of input image regions

Metric: CIDEr and Word IoU

Cornia et al., "Show, Control and Tell: A Framework for Generating Controllable and Grounded Captions", CVPR 2019

Conclusion

- We proposed *the first* **comprehensive** image captioning model that enables **accurate**, **diverse**, **grounded** and **controllable** captioning *at the same time*.
- Our model *outperforms state-of-the-art results* in caption diversity, grounding and controllability, and compares favorably to latest methods in caption quality.

Project Page: http://pages.cs.wisc.edu/~yiwuzhong/Sub-GC.html

Code Repo: https://github.com/YiwuZhong/Sub-GC

Thank you!