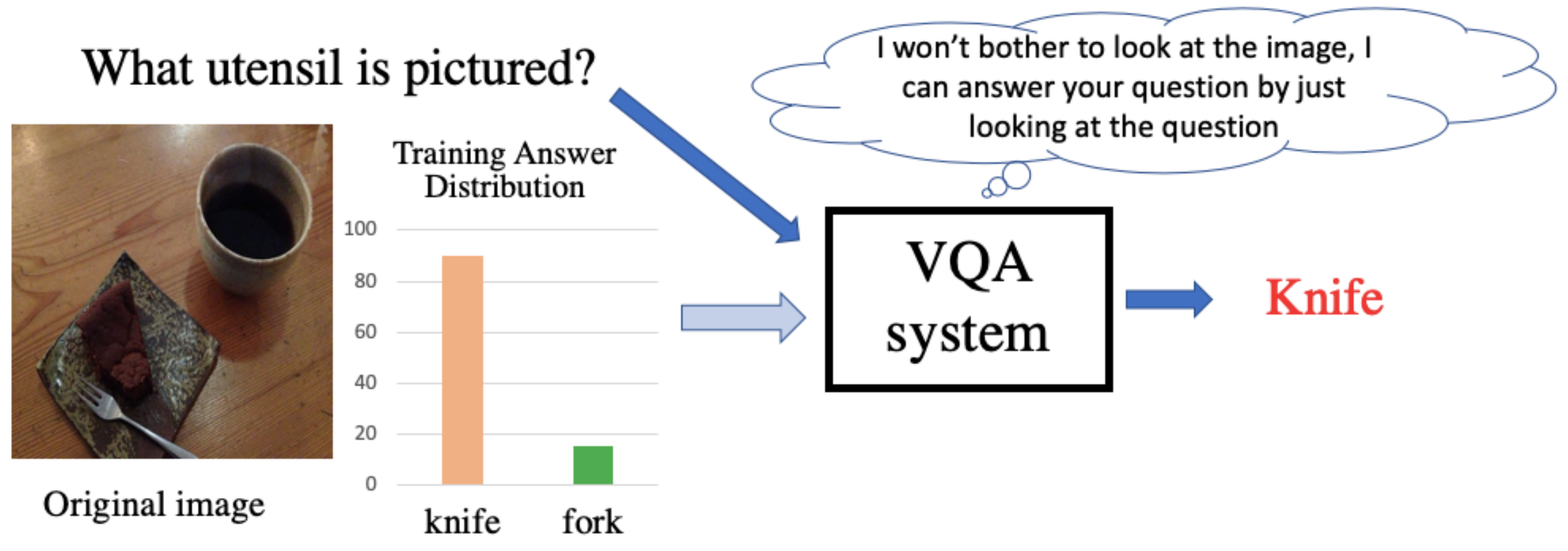
The University of Texas at Austin
Department of Computer Science
College of Natural Sciences

Self-Critical Reasoning for Robust Visual Question Answering

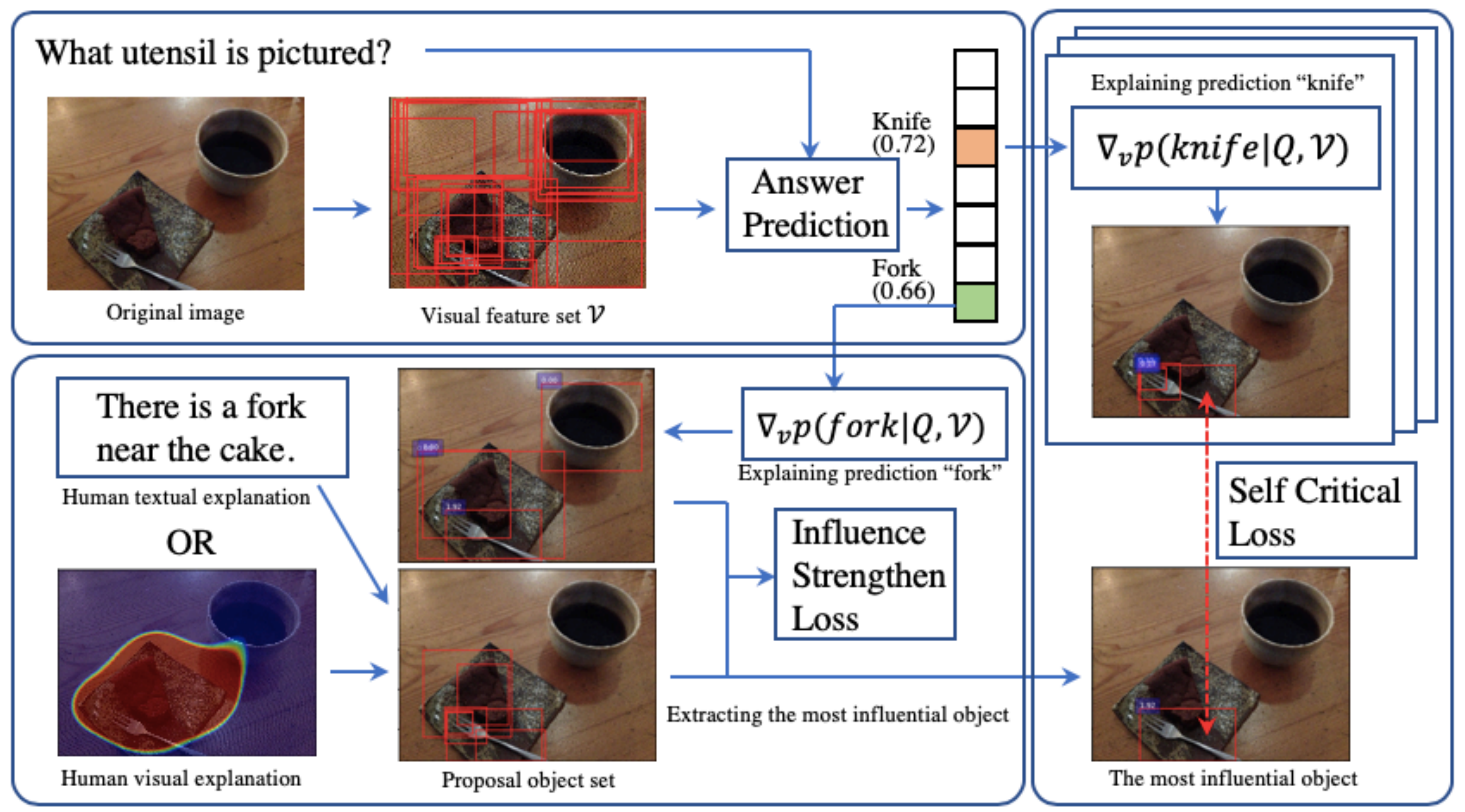
Jialin Wu, and Raymond J. Mooney The University of Texas at Austin

Introduction

- Common VQA systems tend to only capture superficial statistical correlations between QA pairs, especially when training and test set are under different distribution.
- VQA systems should focus on the objects that human would focus.
- We propose two constraints for VQA systems that help the right objects contribute more to the right answers than to the wrong answers.


What utensil is pictured?
I won't bother to look at the image, I can answer your question by just looking at the question
Training Answer Distribution
100
80
60
40
20
0
knife
fork
Original image
VQA system
Knife

Model Overview


What utensil is pictured?
Original image
Visual feature set $\mathcal{V}$
Answer Prediction
Knife (0.72)
Fork (0.66)
Explaining prediction "knife"
$\nabla_v p(knife|Q,\mathcal{V})$
$\nabla_v p(fork|Q,\mathcal{V})$
Explaining prediction "fork"
There is a fork near the cake.
Human textual explanation
OR
Human visual explanation
Proposal object set
Influence Strengthen Loss
Extracting the most influential object
Self Critical Loss
The most influential object

Recognizing and Strengthening Influential Objects

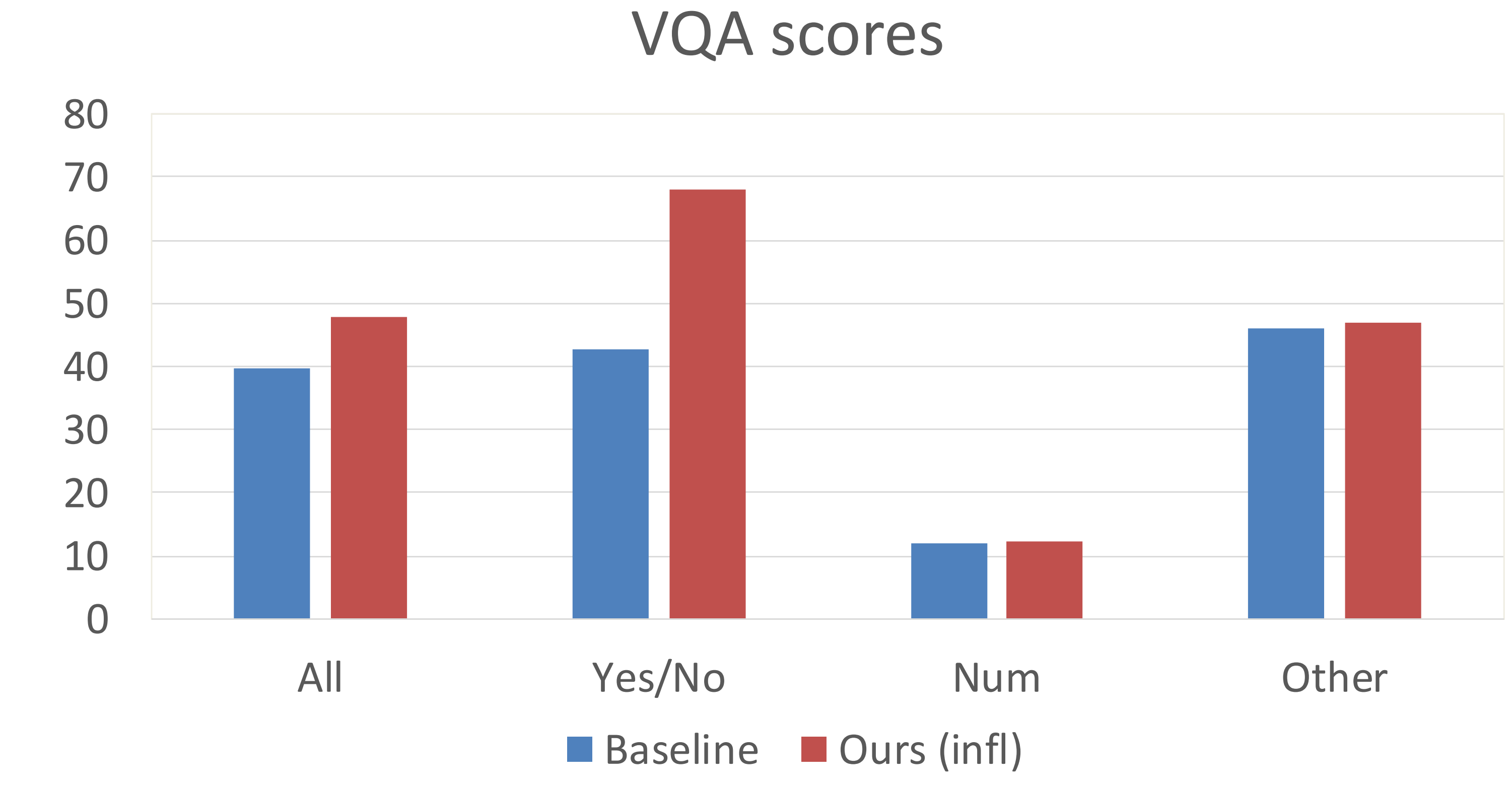
- Extracting a set of influential objects ($\mathcal{I}$) that humans would focus on.
- Enforcing the gradients ($\mathcal{S}(a, \mathbf{v}_i)$) from the correct answer to have the biggest value in at least one of the extracted objects.

$$\mathcal{S}(a, \mathbf{v}_i) := \left(\nabla_{\mathbf{v}_i} P(a|V, q)\right)^T \mathbf{1}$$

$$\mathcal{SV}(a, \mathbf{v}_i, \mathbf{v}_j) = \max\left(\mathcal{S}(a, \mathbf{v}_j) - \mathcal{S}(a, \mathbf{v}_i), 0\right)$$

$$\mathcal{L}_{infl} = \min_{\mathbf{v}_i \in \mathcal{I}} \Big(\sum_{\mathbf{v}_j \in \mathcal{V} \setminus \mathcal{I}} \mathcal{SV}(a_{gt}, \mathbf{v}_i, \mathbf{v}_j) \Big)$$

- Comparing to Up-Down VQA system.



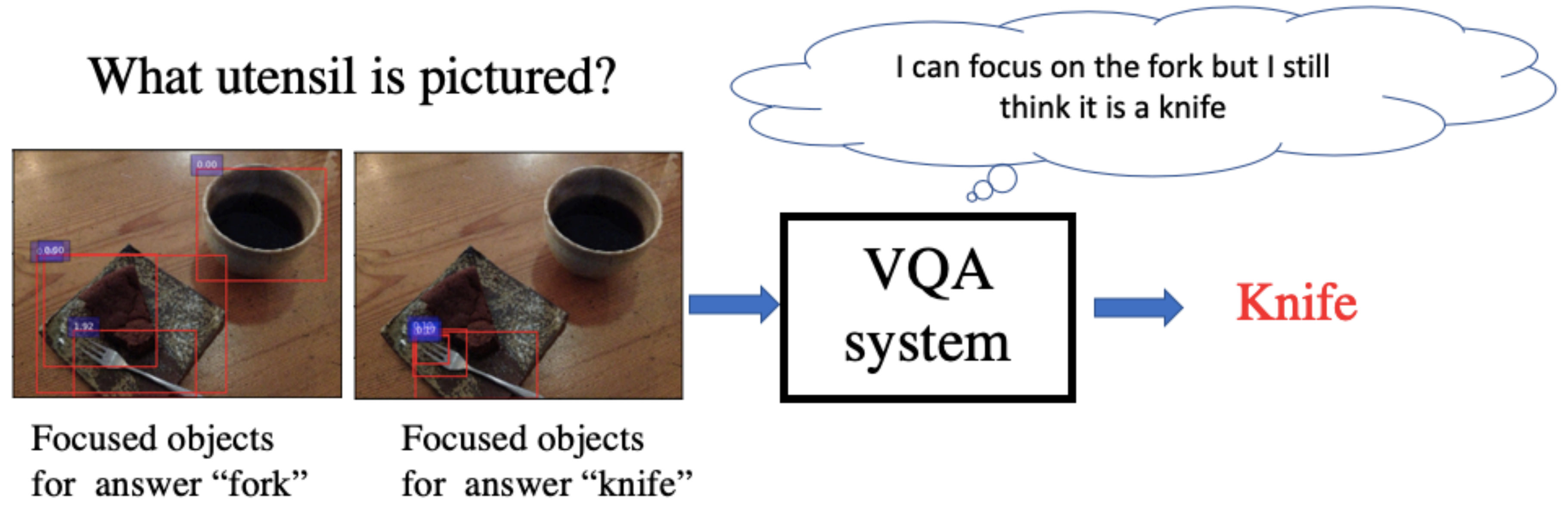
Criticizing Incorrect Dominant Answers

- Although VQA systems focus on the right object for the right answer, but the object could contribute more to the wrong answers.
- Finding the most influential object ($\mathbf{v}^*$) using gradient-based method.

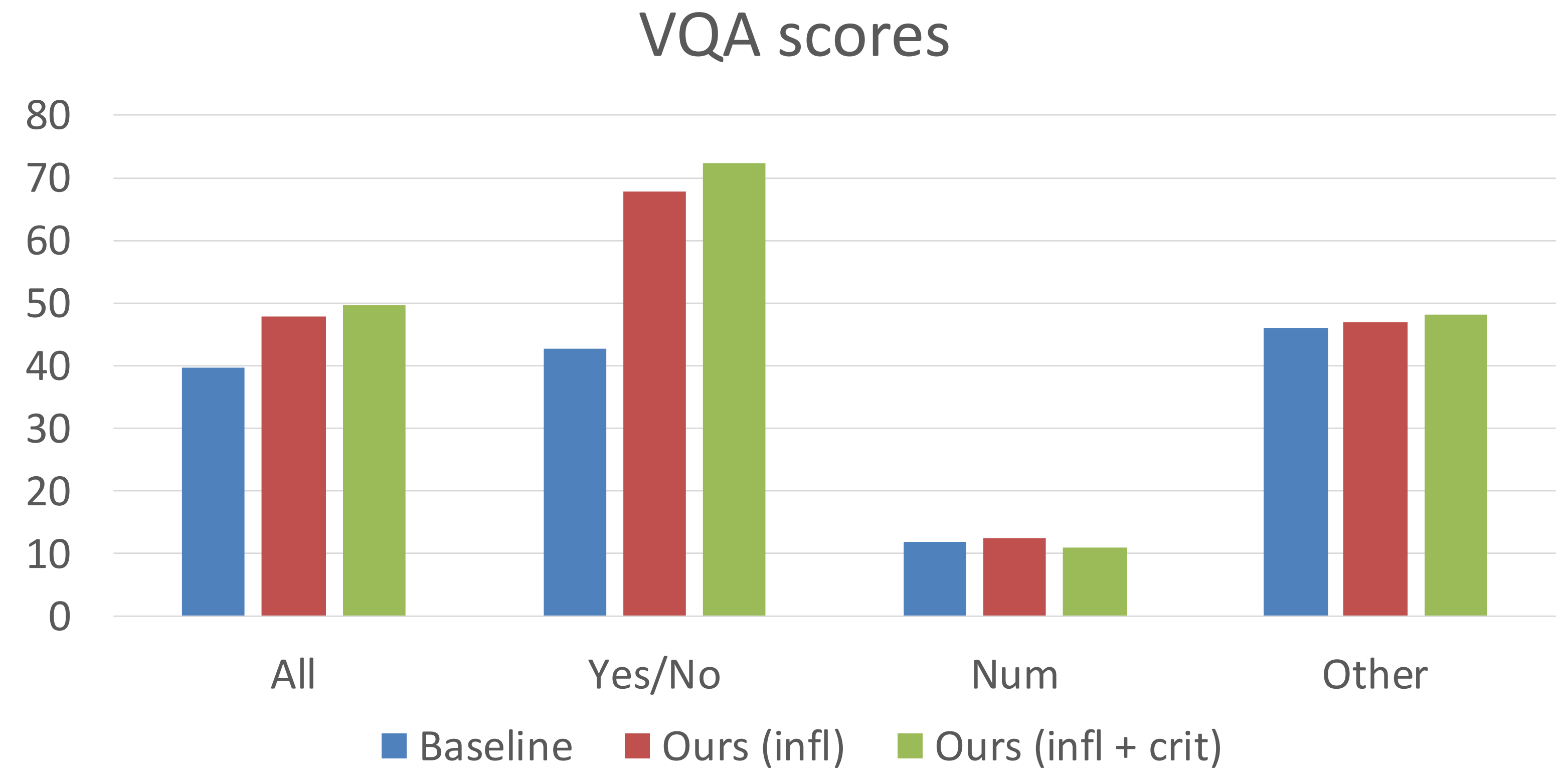
$$\mathbf{v}^* = \arg\min_{\mathbf{v}_i \in \mathcal{I}} \Big(\sum_{\mathbf{v}_j \in \mathcal{V} \setminus \mathcal{I}} \mathcal{SV}(a_{gt}, \mathbf{v}_i, \mathbf{v}_j) \Big)$$

- Enforcing the object to contribute more to the correct answer.

$$\mathcal{L}_{crit} = \sum_{a \in \mathcal{B}} w(a)(\mathcal{S}(a, \mathbf{v}^*) - \mathcal{S}(a_{gt}, \mathbf{v}^*))$$


What utensil is pictured?
I can focus on the fork but I still think it is a knife
Focused objects for answer "fork"
Focused objects for answer "knife"
VQA system
Knife

- Comparing to Up-Down VQA system.



Conclusion

- VQA systems should be able to focus on the right set of objects as human do to predict the right answer
- It is also necessary to prevent the systems from over sensitive to the most common answers.