Answering Questions about Data Visualizations using Efficient Bimodal Fusion (Supplementary Materials)

Kushal Kafle[1] Robik Shrestha[1] Brian Price[2] Scott Cohen[2] Christopher Kanan[1,3,4]
[1]Rochester Institute of Technology [2]Adobe Research [3]Paige [4]Cornell Tech
[1]{kk6055, rss9369, kanan}@rit.edu [2]{bprice, scohen}@adobe.com

In this document, we provide additional results and examples that were omitted from the main paper due to space.

1. Analysis per FigureQA Question Template

Table 1 shows results for PReFIL compared to RN [3, 2] and human baselines [2] for different question templates. The results are from a subset of the Test 2 split in FigureQA. As mentioned in the main document, Test 2 split consists of chart images where the charts have alternated colors compared to the training set, such that the colors are novel for a given chart-type. Test 2 annotations are not publicly available and the results were obtained by sending model predictions to the authors. As seen in table 1, PReFIL outperforms RN for all question templates by a large margin and also outperforms human baseline in 12 out of 15 question templates.

2. More Discussion of Example Outputs

We present additional examples for our PReFIL algorithm for both the DVQA [1] (Fig. 1) and FigureQA (Fig. 2) datasets. For both datasets, we present examples of correct predictions for a variety of examples (top two rows) and some cases of incorrect predictions (bottom row).

For DVQA, PReFIL with oracle OCR is exceedingly capable, with accuracy of over 96% (see main text for details), but it makes some occasional errors. First, since the dynamic encoding is based on the position of words in the chart, PReFIL may detect the wrong word when the words are in close proximity to each other (Fig. 1, bottom left). Second, when the chart elements are partially or fully obscured by the legend, PReFIL often fails to correctly parse the chart data (Fig. 1, bottom center). Finally, for some charts, questions involving multiple measurements are also erroneous, especially when the measurements differ only by a small amount (Fig. 1, bottom right).

For FigureQA, PReFIL again performs well across all categories, surpassing overall human accuracy. PReFIL is capable of answering a wide range of questions across several types of images (Fig. 2, top 2 rows). However, PReFIL often struggles for question template "Is X the smoothest/roughest?" especially for the dot-line style graphs. The errors are more prominent when the legend obscures or intermingles with the chart elements (Fig. 2, bottom left). Since the dots are not connected to each other, it is an extremely difficult task even for attentive human observers. Similarly, PReFIL makes occasional mistakes when comparing elements that are very close to each other (Fig. 2, bottom center and right). However, as seen in Table 1, PReFIL is more accurate than even human observers for comparing two elements.

References

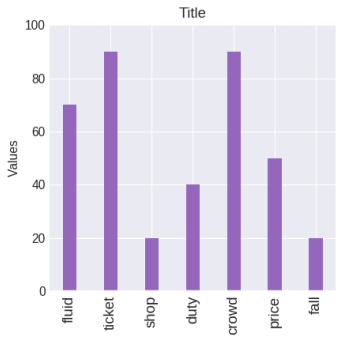
[1] K. Kafle, S. Cohen, B. Price, and C. Kanan. Dvqa: Understanding data visualizations via question answering. In *CVPR*, 2018. 1

[2] S. E. Kahou, A. Atkinson, V. Michalski, A. Kadar, A. Trischler, and Y. Bengio. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*, 2017. 1, 2

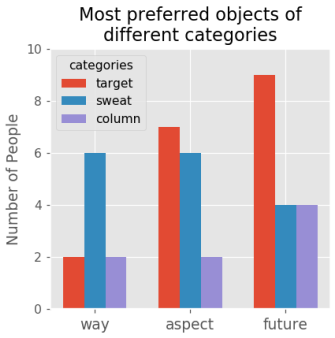
[3] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap. A simple neural network module for relational reasoning. In *NIPS*, 2017. 1, 2

Table 1. Results for PReFIL compared with RN [3, 2] and Human baseline [2] compared with each unique question template in FigureQA.

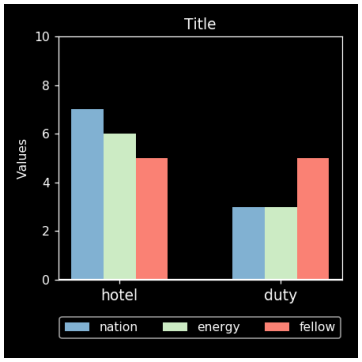
Question Template	**Figure Types**	**RN [3, 2]**	**Human [2]**	**PReFIL (Ours)**
Is X the minimum?	bar, pie	76.78	97.06	**97.20**
Is X the maximum?	bar, pie	83.47	97.18	**98.07**
Is X the low median?	bar, pie	66.69	86.39	**93.07**
Is X the high median?	bar, pie	66.50	86.91	**93.00**
Is X less than Y ?	bar, pie	80.49	96.15	**98.20**
Is X greater than Y ?	bar, pie	81.00	96.15	**98.07**
Does X have the minimum area under the curve?	line	69.57	**94.22**	94.00
Does X have the maximum area under the curve?	line	78.45	95.36	**96.91**
Is X the smoothest?	line	58.57	**78.02**	71.87
Is X the roughest?	line	56.28	**79.52**	74.67
Does X have the lowest value?	line	69.65	90.33	**92.17**
Does X have the highest value?	line	76.23	93.11	**94.83**
Is X less than Y?	line	67.75	90.12	**92.38**
Is X greater than Y?	line	67.12	89.88	**92.00**
Does X intersect Y ?	line	68.75	89.62	**91.25**
Overall	bar,pie,line	72.18	91.21	**92.79**



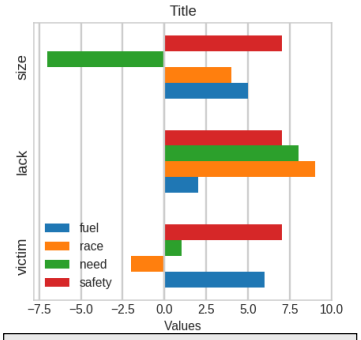
Q. Is each bar a solid color without patterns? A: Yes



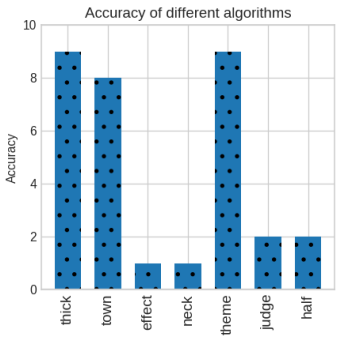
Q: Which object is preferred by the least number of people summed across all the categories? A: way



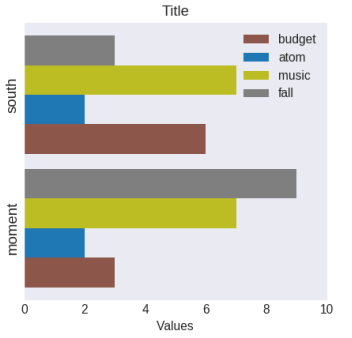
Q. What element does the salmon color represent? A: fellow



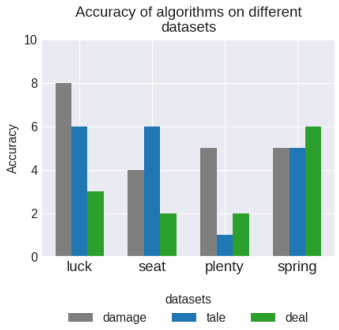
Q. Does the chart contain any negative values? A: Yes



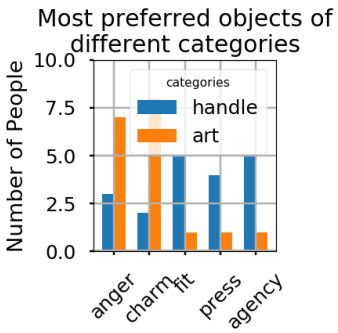
Q: What is the accuracy of the algorithm effect? A: 1



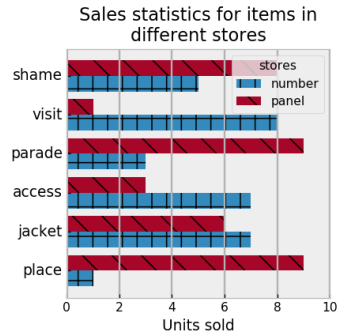
Q. Which group of bars contains the largest valued individual bar in the whole chart? A: moment



Q. What dataset does the steelblue color represent? A: datasets (tale)

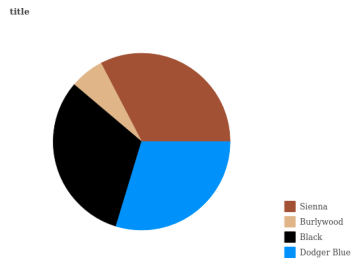


Q: How many people prefer the item art in the category charm? A: 5 (8)

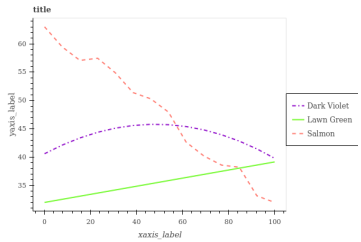


Q. Which item sold the least number of units summed across all the stores? A: place (visit)

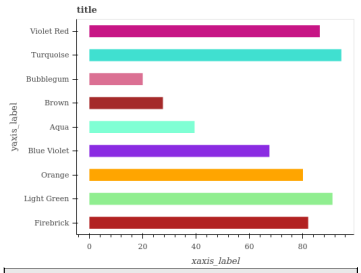
Figure 1. Some example predictions for PReFIL on the DVQA dataset. Red denotes incorrect predictions. For incorrect predictions, correct answer is shown in parenthesis.



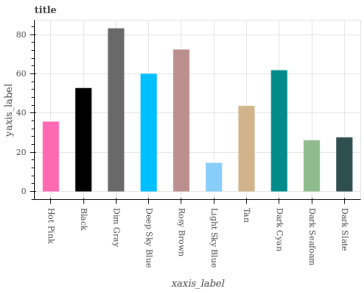
Q. Is Sienna greater than Burlywood? A: Yes



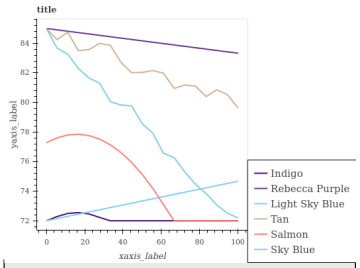
Q: Is salmon the smoothest? A: No



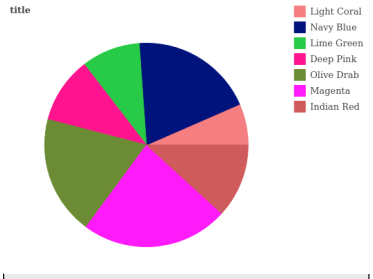
Q. Is Bubblegum greater than Brown? A: No



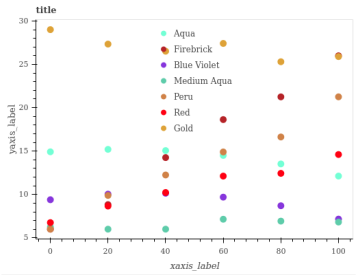
Q. Is Light sky blue the minimum? A: Yes



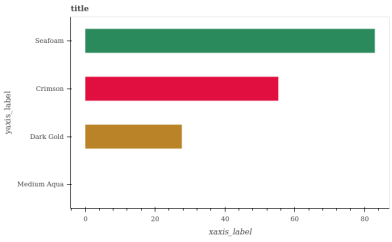
Q: Is Tan the roughest? A: Yes



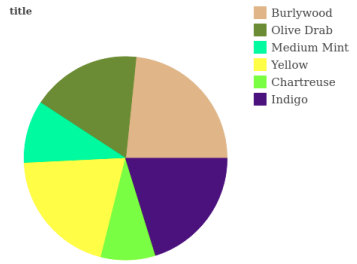
Q.Is Olive Drab the maximum? A: No



Q. Is Red the smoothest? A: Yes (No)



Q: Is Dark Gold the minimum? A: Yes (No)



Q. Is Medium Mint greater than Chartreuse? A: No (Yes)

Figure 2. Some example predictions for PReFIL on the FigureQA dataset. Bottom row shows some incorrect predictions made by PReFIL. Red denotes incorrect predictions. For incorrect predictions, correct answer is shown in parenthesis.