1. Introduction

This document supplements our AUTO-QA dataset described in the main draft. We provide attributes of our dataset in JSON format as shown in figure-1 and we also provide few sample examples of our dataset in JSON format as shown in figure-2. We also include detailed analysis of level-1 and level-2 attention of our various benchmark approaches as shown in figure-3. We provide few sample results, where we predict and visualise our answer for a given set of images and its corresponding question as shown in figure-4 and 5.

Along with this document, we provide our source code and sample JSON in different folders. More details are present on our webpage: https://delta-lab-iitk.github.io/AUTO-QA/

2. Attention Results

We have shown hierarchical attention visualisation results for an image-based model as shown in figure-3. For level-1 attention, these results are obtained using Stack Attention Network, and the level-2 attention is visualised by changing the transparency of images. In this level attention, weights are normalized between zero & one, and transparency of image corresponding to highest normalized level-2 attention weights is set to one, and rest are set to 0.2. We also provide few results of our model, which predict the answer for a given set of images and its question as shown in figure-4 and 5. Our visualisation indicates that where the model is focused on the image for the corresponding question.
Figure 1. This figure shows attributes of our Auto-QA dataset in JSON format. Each sample mainly contains Question type, Question ID, Lidar point information, Video information, programs, Answer information, and split.
Figure 2. This figure shows a few sample examples of our Auto-QA dataset in JSON format.

```
5:
  question_family_index: 0
  question_index: 5
  question: "Which object is the closest towards front left?"
  template_filename: "closest.json"
  lidar_index: 0
  program: [    
  video: "02cf0ce1-699a-373b-86c8-eb6fd5f4697a"
  split: "train"
  answer: "large_vehicle"

6:
  question_family_index: 0
  question_index: 6
  question: "How many large_vehicles are on front left?"
  template_filename: "count.json"
  lidar_index: 0
  program: [    
  video: "02cf0ce1-699a-373b-86c8-eb6fd5f4697a"
  split: "train"
  answer: "1"

7:
  question_family_index: 0
  question_index: 7
  question: "What is the count of vehicles on front right?"
  template_filename: "count.json"
  lidar_index: 0
  program: [    
  video: "02cf0ce1-699a-373b-86c8-eb6fd5f4697a"
  split: "train"
```
Figure 3. This figure shows attention visualization for level-1 (top) and level-2 attention (down) few sample instances of Auto-QA dataset.
| What is the number of vehicles on my front left | Answer: 2 | Which object is closest toward my rear right? | Answer: vehicle | Is there an obstacle toward my rear left? | Answer: false |

Figure 4. This figure shows attention visualisation for sample images of our Auto-QA dataset and its predicted answer for a particular question.
Figure 5. (Few more examples) This figure shows attention visualisation for sample images of our Auto-QA dataset and its predicted answer for a particular question.