Object Referring in Videos with Language and Human Gaze- Supplementary Material

We provide the following additional materials:

- **Annotation interface**
- **Gaze Feature encoding**
- **More qualitative results**
- **Accompany video for interface and video examples**

1. Annotation Interface

We annotate a dataset of textual attributes and descriptions for objects in videos. We collected the annotations on Cityscapes dataset [1]. We crowdsourced the annotation task on Amazon Mechanical Turk (AMT). Firstly, we show the set of instructions as in Fig. 5 to the workers. Here, we try to convey a few Do's and Dont's for writing descriptions for the objects along with a visual example. As described in the paper, the goal of the annotation is for cooperative conversations. That is, the descriptions need to be truthful, informative, relevant, and brief for co-observers to find the target objects easily and unambiguously.

Referring Expression annotation. The interface that we show to AMT workers is shown in Fig. 2. Worker is obliged to see the video in the beginning. We freeze all the fields of the interface and it is opened only if the video is viewed at least once. The video is embedded on top of panel of the interface. The video stops with the final frame of the video which is allowed to be annotated as shown in Fig. 2. We display all the previous annotations of the objects in the final frame to avoid repeated object annotations. We provide a Play Video button and encourage the Workers to re-watch the video for better scene understanding and object description.

The worker is then asked to draw a tight bounding box on an unannotated object. For the corresponding annotated objects, he/she fills the fields of object "Attributes" (*e.g.* class name, color, and size) and the "Relationship with the observer (*e.g.* in front, on the right side, and far away)". Finally, the Workers need to describe the object in a short sentence that help people to identify the object uniquely among all the objects. This annotation for attributes and relationships **help** significantly; it makes sure that the Workers have made enough effort to understand the object and the scene well. The attributes and relationships will be explored for referring expression generation in our future work.

We display the list of objects annotated by the worker on the left under the heading "Box list" and we provide an option to update/remove a particular object from the list. A Worker is allowed to annotate two objects per video at maximum for the diversity of annotations. In Fig. 3, we have shown an example of annotation of three objects with their corresponding descriptions for a particular video. Fig. 4 shows a sample annotation of Attributes and Relationships for the objects annotated in Fig. 3. Please see the accompany video for video examples and kindly refer to Fig. 2 for a better view of the interface.

Gaze recording. Fig. 6 demonstrates the web interface that we show to workers in AMT. Since the recording of gaze involves recording of the faces, we ensure that workers read the privacy policies before they start the recording task. This primarily includes that their faces will be recorded during the tasks and the recordings will be used for research purposes only. Our aim is to record all the gaze for bounding box annotations of the objects from the task of referring expression. We have explained the annotation process in the main paper. Here, we have also shown some examples of recorded frontal view of the annotators for gaze recording in Fig. 1.

2. Gaze feature encoding

Workers annotate by gazing at the corners of the image displayed on the canvas. These corner gazes are used to find the mapping function from camera coordinates to image coordinates. This mapping function is later used to compute the gaze features. Given the corners in image coordinates I_1, I_2, I_3, I_4 of gazes G_1, G_2, G_3, and G_4. We compute the gaze features (F) by the following procedure as shown in Algorithm. 1.

Mapping function. We find the mapping function to map the coordinates from screen/camera coordinates to image coordinates. We get the system of linear equations from the four corners which we annotated for the calibration as

```
Data: Gaze of corners $G_1, G_2, G_3, G_4$,
      Corners(image coordinates) $I_1, I_2, I_3, I_4$ and
      Bounding boxes $BB$
Result: Gaze Features($F$)
Initialization: Corners (camera coordinates) $C_1, C_2,$
  $C_3, C_4$;
Begin:
$C_i = DAN(G_i)$ [3]; i=1,2,3,4.;
$I_i = A * C_i + B$ (as in Eq. 1)where $A = (a1, a2)$ and
  $B = (b1, b2)$;
Solve for A and B (See Eq. 2);
while For all Gaze $g_i$ and bounding box $BB_i$ do
    $p_i = A * DAN(g_i) + B$;
    $heatmap = gaussian\_2d(Image, p_i, \sigma = 20)$;
    if $p_i$ within image then
        $F_i = mean(heatmap[BB_i])$;
    else
        $BB\_norm =$
          $heatmap[BB_i]/max(heatmap[BB_i])$
          $F_i = mean(BB\_norm)$;
    end
end
Gaze Features= F;
End;
```

Algorithm 1: Extracting Gaze Features

shown in Algorithm. 1. Given the system of linear equations with more number of equations than the variables, we have overdetermined equation

$$y = Hx \tag{1}$$

The solution x can be derived by minimizing the energy of error:

$$J = ||y - Hx||_2^2$$
$$J = (y - Hx)^T(y - Hx)$$
$$J = y^Ty - 2y^THx + x^TH^THx$$

Making derivative of J to zero, we get,

$$x = (H^TH)^{-1}H^Ty, \tag{2}$$

assuming H^TH is invertible.

3. Qualitative Results

In this section, we present more qualitative results on Cityscapes dataset from different models to show the advantage of using multiple modalities: NLOR [2], Ours:(I,D,O) and Ours:(I,D,O,G) in Fig. 7. From column 4, we see how gaze improve the object localization results over already given depth information, motion information and the temporal-spacial context information.

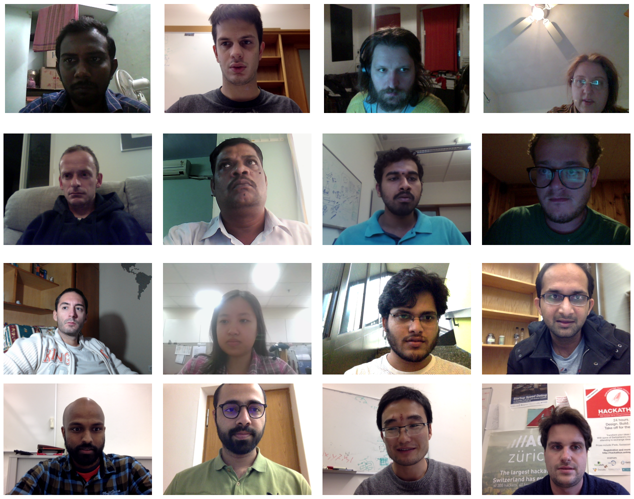


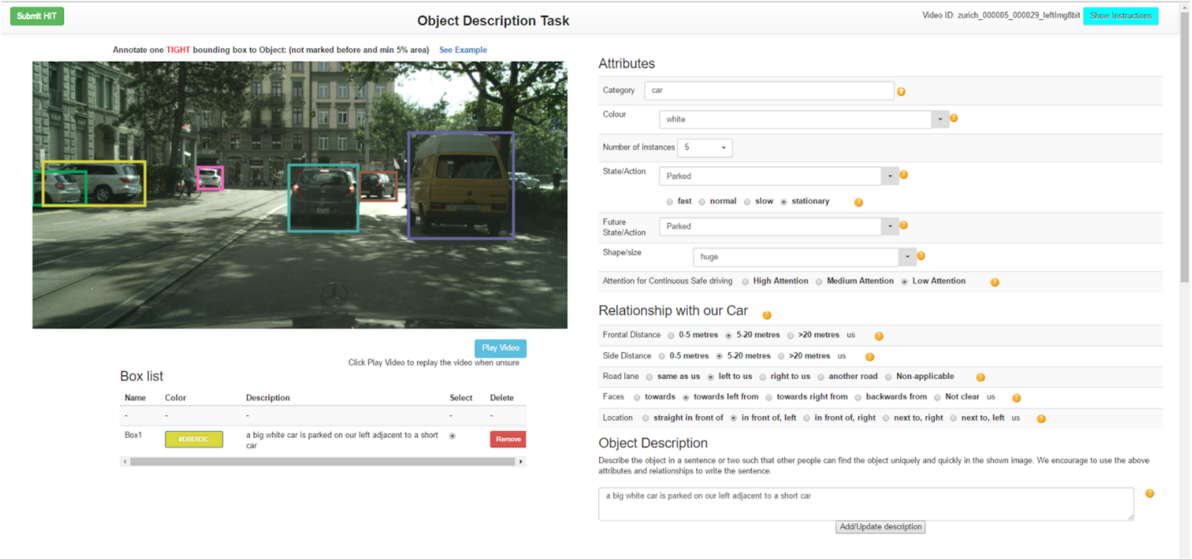
Figure 1: Selected images of gaze recording from our annotations.

References

[1] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[2] R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, and T. Darrell. Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4555–4564, 2016. 2

[3] M. Kowalski, J. Naruniec, and T. Trzcinski. Deep alignment network: A convolutional neural network for robust face alignment. *arXiv preprint arXiv:1706.01789*, 2017. 2



Submit HIT

Object Description Task

Video ID: zurich_000005_000029_leftImg8bit

Show Instructions

Annotate one TIGHT bounding box to Object: (not marked before and min 5% area) See Example

Play Video

Click Play Video to replay the video when unsure

Box list

Name	Color	Description	Select	Delete
-	-	-	-	-
Box1	#DBD83C	a big white car is parked on our left adjacent to a short car	●	Remove

Attributes

Category: car

Colour: white

Number of instances: 5

State/Action: Parked

○ fast ○ normal ○ slow ● stationary

Future State/Action: Parked

Shape/size: huge

Attention for Continuous Safe driving ○ High Attention ○ Medium Attention ● Low Attention

Relationship with our Car

Frontal Distance ○ 0-5 metres ● 5-20 metres ○ >20 metres us

Side Distance ○ 0-5 metres ● 5-20 metres ○ >20 metres us

Road lane ○ same as us ● left to us ○ right to us ○ another road ○ Non-applicable

Faces ○ towards ● towards left from ○ towards right from ○ backwards from ○ Not clear us

Location ○ straight in front of ● in front of, left ○ in front of, right ○ next to, right ○ next to, left us

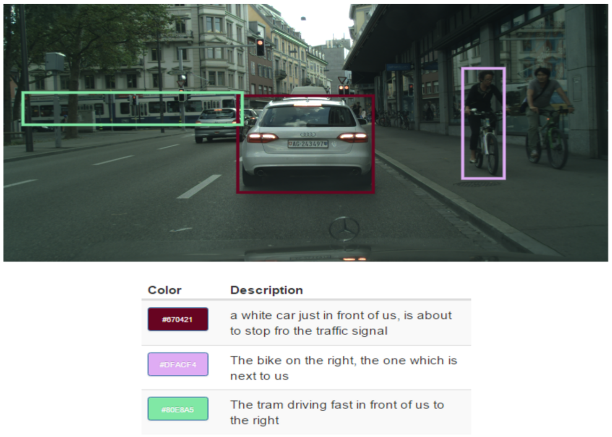
Object Description

Describe the object in a sentence or two such that other people can find the object uniquely and quickly in the shown image. We encourage to use the above attributes and relationships to write the sentence.

a big white car is parked on our left adjacent to a short car

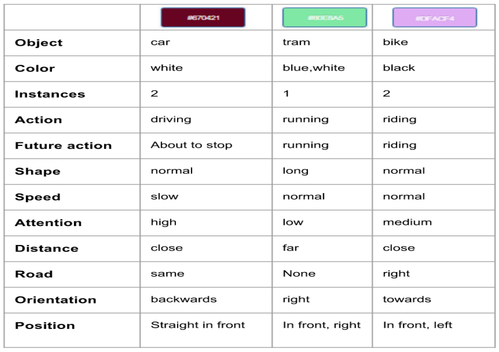
Add/Update description

Figure 2: Snapshot of the annotation interface that we show to the workers for Object Description task.



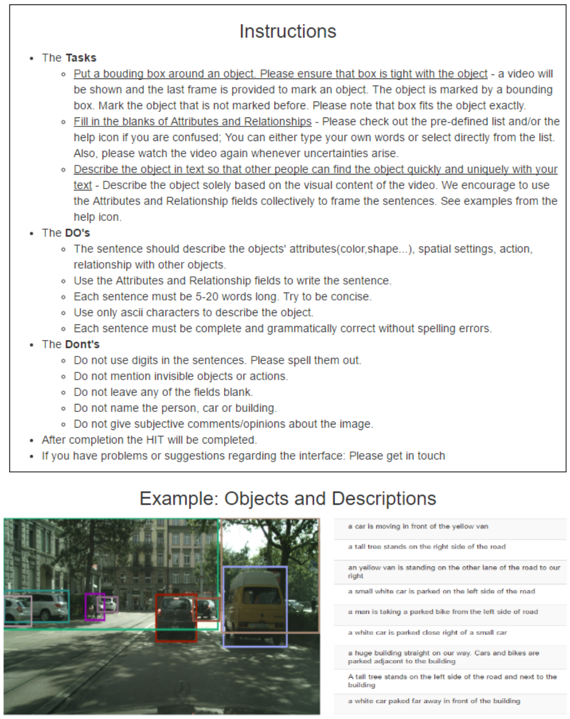
Color	Description
#670421	a white car just in front of us, is about to stop fro the traffic signal
#DFACF4	The bike on the right, the one which is next to us
#80E8A5	The tram driving fast in front of us to the right

Figure 3: Sample annotations of objects and their descriptions.



	#670421	#80E8A5	#DFACF4
Object	car	tram	bike
Color	white	blue,white	black
Instances	2	1	2
Action	driving	running	riding
Future action	About to stop	running	riding
Shape	normal	long	normal
Speed	slow	normal	normal
Attention	high	low	medium
Distance	close	far	close
Road	same	None	right
Orientation	backwards	right	towards
Position	Straight in front	In front, right	In front, left

Figure 4: Sample annotation of object attributes.



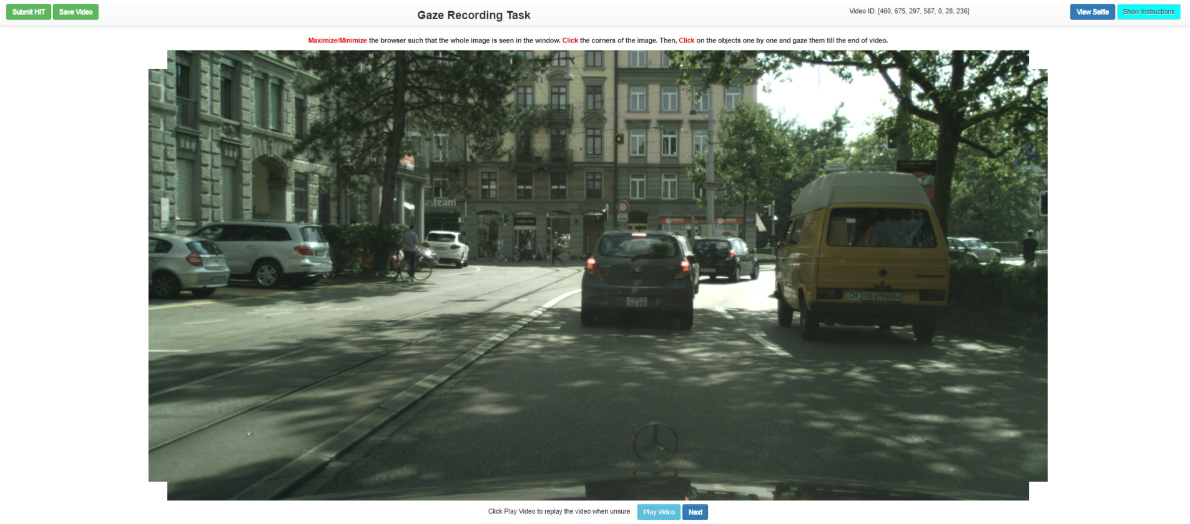
Instructions

- The **Tasks**
  - Put a bouding box around an object. Please ensure that box is tight with the object - a video will be shown and the last frame is provided to mark an object. The object is marked by a bounding box. Mark the object that is not marked before. Please note that box fits the object exactly.
  - Fill in the blanks of Attributes and Relationships - Please check out the pre-defined list and/or the help icon if you are confused; You can either type your own words or select directly from the list. Also, please watch the video again whenever uncertainties arise.
  - Describe the object in text so that other people can find the object quickly and uniquely with your text - Describe the object solely based on the visual content of the video. We encourage to use the Attributes and Relationship fields collectively to frame the sentences. See examples from the help icon.
- The **DO's**
  - The sentence should describe the objects' attributes(color,shape...), spatial settings, action, relationship with other objects.
  - Use the Attributes and Relationship fields to write the sentence.
  - Each sentence must be 5-20 words long. Try to be concise.
  - Use only ascii characters to describe the object.
  - Each sentence must be complete and grammatically correct without spelling errors.
- The **Dont's**
  - Do not use digits in the sentences. Please spell them out.
  - Do not mention invisible objects or actions.
  - Do not leave any of the fields blank.
  - Do not name the person, car or building.
  - Do not give subjective comments/opinions about the image.
- After completion the HIT will be completed.
- If you have problems or suggestions regarding the interface: Please get in touch

Example: Objects and Descriptions

- a car is moving in front of the yellow van
- a tall tree stands on the right side of the road
- an yellow van is standing on the other lane of the road to our right
- a small white car is parked on the left side of the road
- a man is taking a parked bike from the left side of road
- a white car is parked close right of a small car
- a huge building straight on our way. Cars and bikes are parked adjacent to the building
- A tall tree stands on the left side of the road and next to the building
- a white car paked far away in front of the building

Figure 5: The set of instructions shown to the workers. An example of annotation is also shown on the bottom.



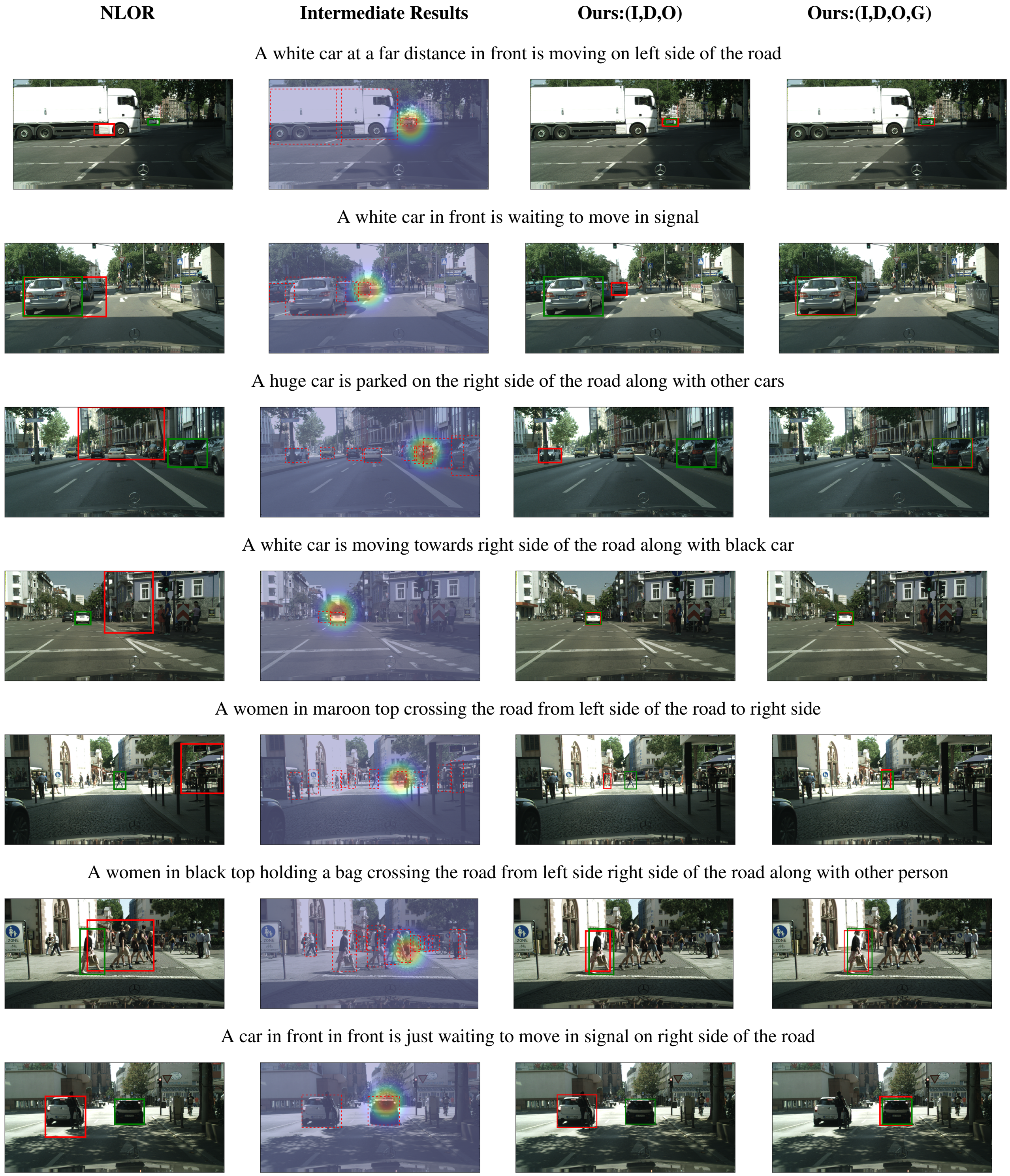
Gaze Recording Task

Video ID: [460, 675, 297, 587, 0, 28, 236]

Maximize/Minimize the browser such that the whole image is seen in the window. Click the corners of the image. Then, Click on the objects one by one and gaze them till the end of video.

Click Play Video to replay the video when unsure

Figure 6: Snapshot of the annotation interface that we show to the workers for the task of Gaze recording.



A white car at a far distance in front is moving on left side of the road

A white car in front is waiting to move in signal

A huge car is parked on the right side of the road along with other cars

A white car is moving towards right side of the road along with black car

A women in maroon top crossing the road from left side of the road to right side

A women in black top holding a bag crossing the road from left side right side of the road along with other person

A car in front in front is just waiting to move in signal on right side of the road

Figure 7: Some qualitative results: NLOR, Ours(I,D,O) and Ours:(I,D,O,G). These results are obtained on the Cityscapes dataset. Green boxes represent ground truth box and Red boxes represent the predicted box. We mainly show how gaze helps in improving the results.