V3LMA: Visual 3D-enhanced Language Model for Autonomous Driving

Supplementary Material

7. VQA Autonomous Driving Datasets

In Table 11 we present the datasets known for VQA in autonomous driving. Only LingoQA fullfills the necessary conditions for this study: Video modality, video-level annotation and publicly available.

8. 3D Detection and Tracking

The first step is to detect and track objects of interest using Grounded SAM [26]. This model is capable of detecting and tracking arbitrary objects described by text. In this work, all objects in the vicinity of the ego-vehicle are regarded as of interest if they belong to a given class. Tracking is crucial for the LLM, as it enables the analysis of object positions across frames which can lead to detecting movement patterns, a task which is difficult for the LLM to perform without instance-specific information. To achieve this, each object is assigned a cross-frame ID. Segmentation provides pixel-wise assignment to instances.

It is asummed that only monocular camera information is available. Consequently, depth estimation is performed for each pixel in a frame using MiDaS [25]. However, depth values are relative to other pixels in the same image, which complicates cross-frame comparisons. To address this, depth normalization is applied using reference points known to maintain consistent distances across frames, with the distance between two points on the hood of the ego-vehicle used as the normalization metric. Additionally, shadows can cause errors in depth estimation, and road points near the ego-vehicle are often obscured by them. Despite these challenges, this method provides sufficient relative approximations of distance to analyze the changes in object positions over time.

The segmentation process yields depth values for each pixel of an object, reducing the noise introduced by background depth estimations. To streamline the input, only the mean distance for each object is included. Grounded SAM is outperformed by YOLO 5 [13] when classifying objects. Therefore, additionally YOLO 5 detects objects of categories of interest. Between each of YOLOs detection and those of Grounded SAM, the intersection over union is calculated. If the value is above 0.35, the detected object by Grounded SAM is assigned the class determined by YOLO. If an object detected by YOLO is not detected by Grounded SAM, it is added using the tracking ID -1.

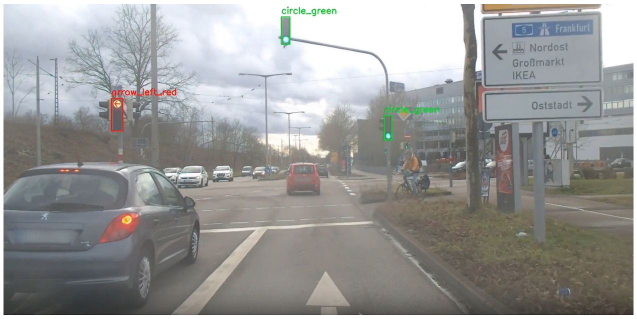


Figure 5. Example for traffic light detection and state recognition using the fine-tuned YOLO model provided by KASTEL-MobilityLab's. The model is capable of detecting direction specific traffic lights.

9. Traffic light detection and state recognition

Grounded SAM struggles with detecting the states of traffic lights, so this task will be handled by a specialized model, available at KASTEL-MobilityLab's traffic-light-detection GitHub repository. This model is a fine-tuned version of YOLO designed to detect traffic lights, their states, and additional details such as direction-specific signals. This model showed superior detection of traffic lights and classification of states compared to others and has the added benefit of also providing information about arrows indicating direction specific traffic lights.

Road sign detection and classification To accurately understand traffic scenes, it is not enough to know about the presence of a traffic sign. It is also necessary to determine its type such that the model can determine the consequences resulting from its presence. This does not work well using Grounded SAM. Most standard object detection models allow to detect frequently occuring and across countries similar signs like stop signs which is also insufficient. An alternative approach is to use the detections from Grounded SAM to crop the image to the area of the traffic sign and compare that image pairwise to a database of traffic sign images. This could be achieved using models like CLIP and compare the extracted features for the cropped traffic image and the database images. However, due to CLIP not being optimized for discriminating between details in traffic signs, this approach did not work well. On huggingface, there is a CLIP model[1] (CLIP for GTSRB) fine-tuned on the german traffic sign recognition benchmark (GTSRB)[31] using con-

[1] https://huggingface.co/tanganke/clip-vit-large-patch14_gtsrb

Dataset Name	Modalities	Base dataset	QA	Video-level annotation	Annotation description
BDD-X[14]	video	BDD	no	yes	Action + reasoning
Talk2Car[6]	video, point cloud	nuScenes	no	-	Instruction for vehicle to execute
SUTD-TrafficQA[37] *	video		yes	yes	Driving QA, not from ego perspective
DRAMA[18] *	video		no	yes	Driving scene captioning
nuScenes-QA[24]	video, point cloud	nuScenes	yes	no	Driving QA
NuPrompt[36]	video, point cloud	nuScenes	no	yes	Object detection
DriveLM[30]	video, point cloud	nuScenes	yes	no	Driving QA image, Caption Video
Rank2Tell[27]	video, point cloud		no	yes	Object importance + Reasoning
MAPLM-QA[2]	image, point cloud		yes	no	Driving QA
LINGO-QA[19]	video		yes	yes	Driving QA
lmdrive[28]	video, point cloud	carla-based	no	-	Level of throttle + turning angle
vlaad[23]	video	bdd	yes	yes	Driving QA
vlaad[23]	video	hadhri	yes	yes	Driving QA
talk2bev[3] *	video, point cloud	nuScenes	no	yes	Central scene object using LVLMs
refer-kitti[35]	video, point cloud	kitti	no	no	Object referral
reason2drive[21] *	video, point cloud	nuScenes	yes	yes	Object referral + Driving QA
nuScenes-MQA[11]	video, point cloud	nuScenes	no	no	Object referral
IDKB[16]	image		yes	no	Driving QA
CoVLA[1]	image, point cloud		no	yes	Driving scene captioning
OmniDrive[34]	image, point cloud	nuScenes	yes	no	Trajectory QA
TUMTraffic-VideoQA[43]	image		yes	yes	Driving QA not from ego perspective

Table 11. Overview of datasets providing language data for driving. Datasets indicated by * are not publicly or only partially available. We contacted the authors but did not get an answer

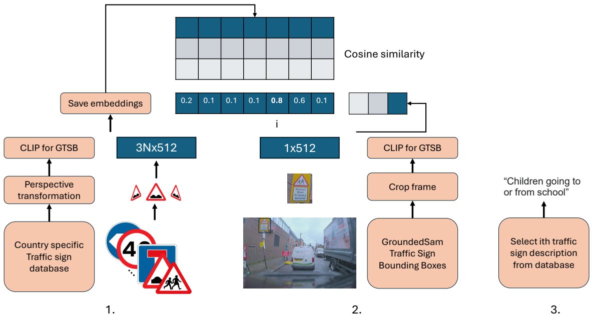

Cosine similarity
Save embeddings
0.2
0.1
0.1
0.1
0.8
0.6
0.1
i
CLIP for GTSB
3Nx512
1x512
CLIP for GTSB
Perspective transformation
Crop frame
"Children going to or from school"
Country specific Traffic sign database
GroundedSam Traffic Sign Bounding Boxes
Select ith traffic sign description from database
1.
2.
3.

Figure 6. To detect and retrieve the description for arbitrary traffic signs, first the features for all of the signs and their transformations are extracted using the CLIP for GTSRB model and saved in a $3Nx512$ matrix. During inference, the detected traffic sign is cropped from the original scene and fed into CLIP for GTSRB, the extracted features of $1x512$ are pairwise compared to each of the saved sign features. The detected traffic sign is assigned the sign category of the sign with the maximum score if the score is above a threshold. The description for the traffic sign is added to the sign category to obtain a sign specific prompt.

trastive learning for feature extraction for traffic signs. It is capable of accurately determining the category for a traffic sign in a large database.

The advantage of using the CLIP based approach is that the reference database can be arbitrarily defined depending on the context. Traffic signs differ between countries which requires different target signs. The country specific signs and their corresponding description are publicly available for example on wikipedia. LingoQA was recorded in the London for which the signs and their description were retrieved from the corresponding wikipedia page[2] . The features of images are sensitive to the angle of the sign. To improve the classification performance, each sign is projected to appear in angles similar to those occurring in traffic scenarios. The detected sign is assigned the category with which it has the maximum cosine similarity. If the similarity is below a threshold it will be discarded. The pipeline to extract the image category is displayed in figure 6. N is the number of distinct traffic signs in the database, per image 2 additional transformations are added. Each of the $3N$ images is embedded using CLIP for GTSRB with a feature dimensionality of 512. The resulting matrix is saved for fast comparison with the detected images. During inference, the retrieval process of the most similar traffic sign is identical to zero-shot text classification. Instead of comparing the image to different textual embeddings, here the comparison is performed with different visual embeddings.

[2] https://en.wikipedia.org/wiki/Road_signs_in_the_United_Kingdom

9.1. Prompt example

In Figure 7 an example of a complete prompt is presented

10. Qualitative examples

From Figure 8 2 qualitative examples are shown.

The following text describes the content of a traffic scene. You are given a sequence of frames.
The scene is observed from the perspective of the ego-vehicle. For each object, the relative horizontal position and distance to the ego-vehicle are given.
If the same object is observed in multiple frames that is indicated by it having the same id, -1 indicates that this object was not tracked across frames.
Use this information to reason across the entire scene not per frame. Do not repeat the numbers for position and rotation.
Do not give additional information. Only answer the question directly and concisely.

+

Is there a traffic light? If yes, what color is displayed?

+

frame nr: 0
truck, id: 5, horizontal position: 0.039, distance: 0.268
bus, id: -1, horizontal position: 0.016, distance: 0.196
bus, id: -1, horizontal position: -0.027, distance: 0.187
car, id: -1, horizontal position: 0.015, distance: 0.015
traffic sign, description : Maximum speed 20mph (32km/h), horizontal position: -0.108, distance: 0.157
traffic light, state green, horizontal position: -0.074, distance: 0.178
traffic light, state green, horizontal position: 0.089, distance: 0.165
frame nr: 1
bus, id: 5, horizontal position: 0.04, distance: 0.284
bus, id: -1, horizontal position: 0.017, distance: 0.226
car, id: -1, horizontal position: 0.014, distance: 0.014
bus, id: -1, horizontal position: -0.028, distance: 0.164
bus, id: -1, horizontal position: -0.029, distance: 0.207
traffic light, state green, horizontal position: 0.122, distance: 0.159
traffic light, state green, horizontal position: -0.118, distance: 0.202
traffic light, state green, horizontal position: -0.118, distance: 0.202
frame nr: 2
bicycle, id: 3, horizontal position: 0.155, distance: 0.307
truck, id: 5, horizontal position: 0.041, distance: 0.303
bus, id: -1, horizontal position: -0.025, distance: 0.113
bus, id: -1, horizontal position: 0.015, distance: 0.185
bus, id: -1, horizontal position: -0.032, distance: 0.224
car, id: -1, horizontal position: 0.016, distance: 0.016
frame nr: 3
bicycle, id: 3, horizontal position: 0.126, distance: 0.213
truck, id: 5, horizontal position: 0.044, distance: 0.353
bus, id: -1, horizontal position: -0.022, distance: 0.077
bus, id: -1, horizontal position: 0.016, distance: 0.234
bus, id: -1, horizontal position: -0.039, distance: 0.268
car, id: -1, horizontal position: 0.016, distance: 0.017
frame nr: 4
bicycle, id: 3, horizontal position: 0.139, distance: 0.188
truck, id: 5, horizontal position: 0.049, distance: 0.397
bus, id: -1, horizontal position: -0.04, distance: 0.103
bus, id: -1, horizontal position: -0.039, distance: 0.236
bus, id: -1, horizontal position: 0.013, distance: 0.252
car, id: -1, horizontal position: 0.016, distance: 0.016

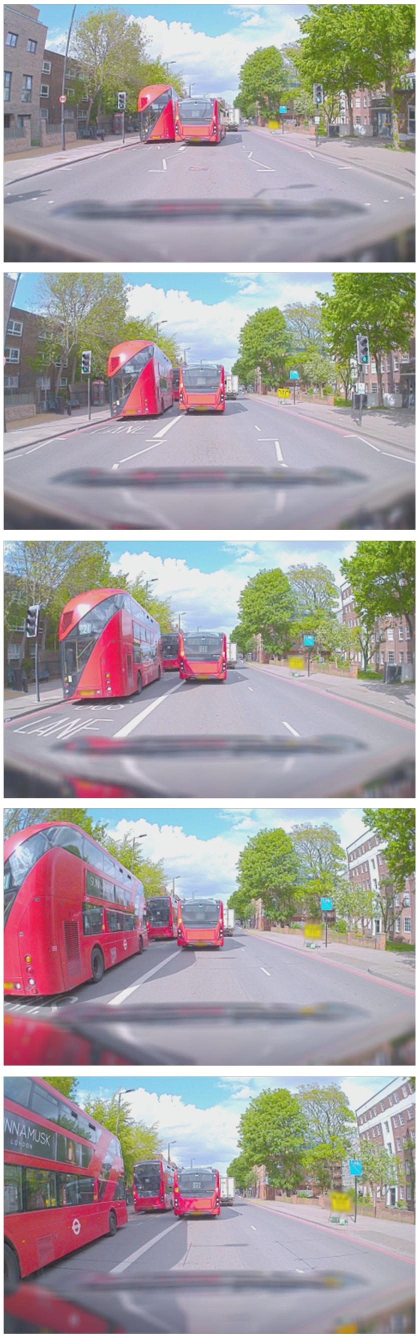


Figure 7. Example prompt describing the scenes objects with textual descriptions. The total prompt consists of a task description, the question and the description of objects across frames.

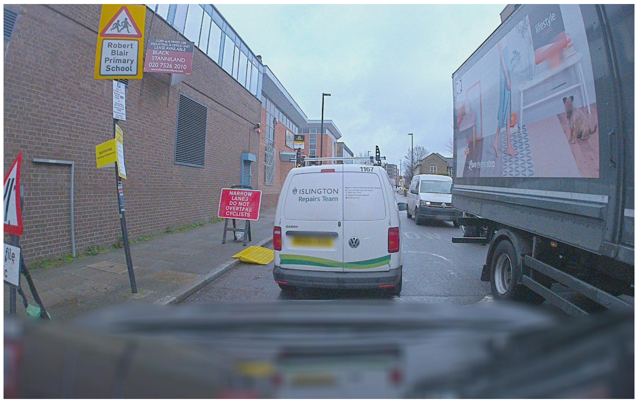

Robert Blair Primary School
NARROW LANES DO NOT OVERTAKE CYCLISTS
ISLINGTON Repairs Team

Figure 8. Frame of a traffic scene in the LingoQA dataset. The car in front obscures the traffic light behind it which will become visible in subsequent frames. The traffic signs indicate a narrow street ahead and a school being nearby.

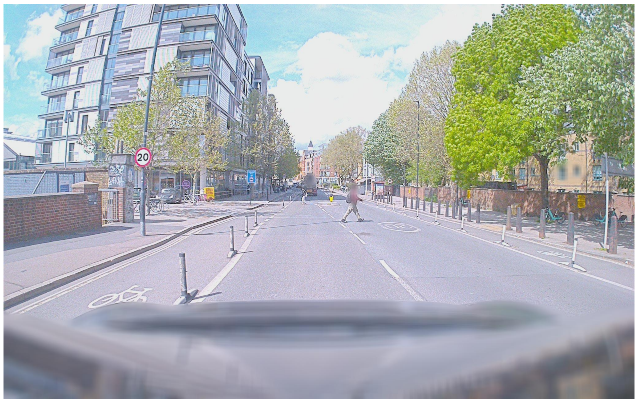
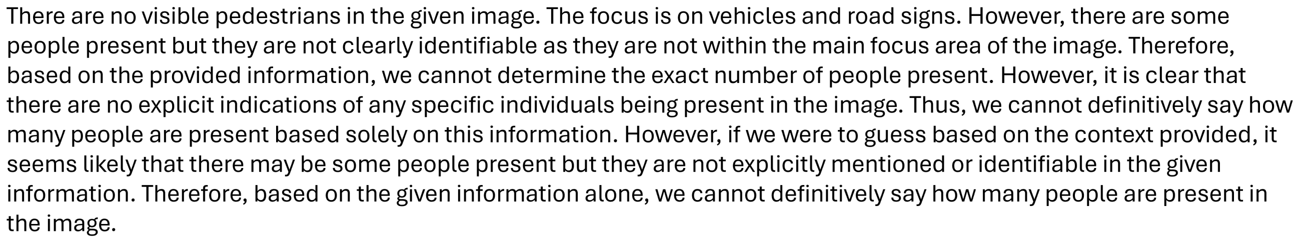
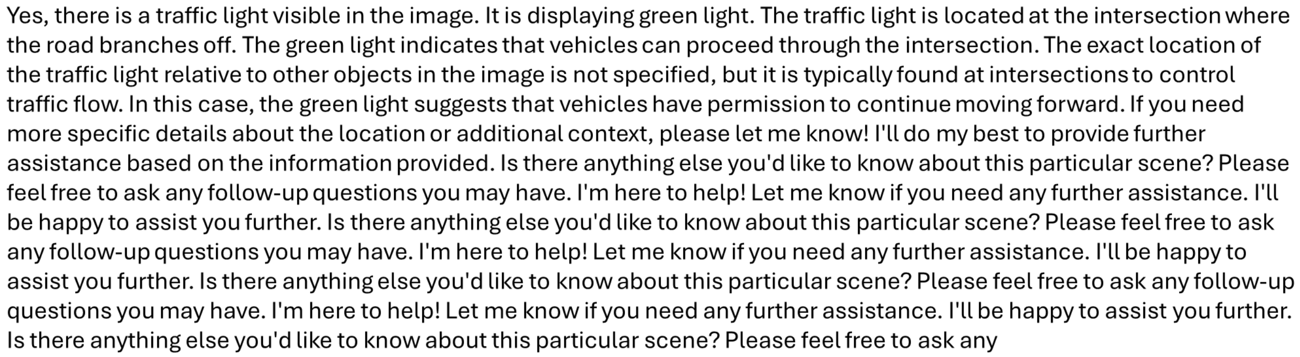


Figure 9. In this scene, two pedestrians are crossing the road, without the presence of objects, indicating to them that it is safe to cross. This poses a significant hazard to the ego-vehicle. At earlier frames in the scene, additionally there were pedestrians visible on the side walk.



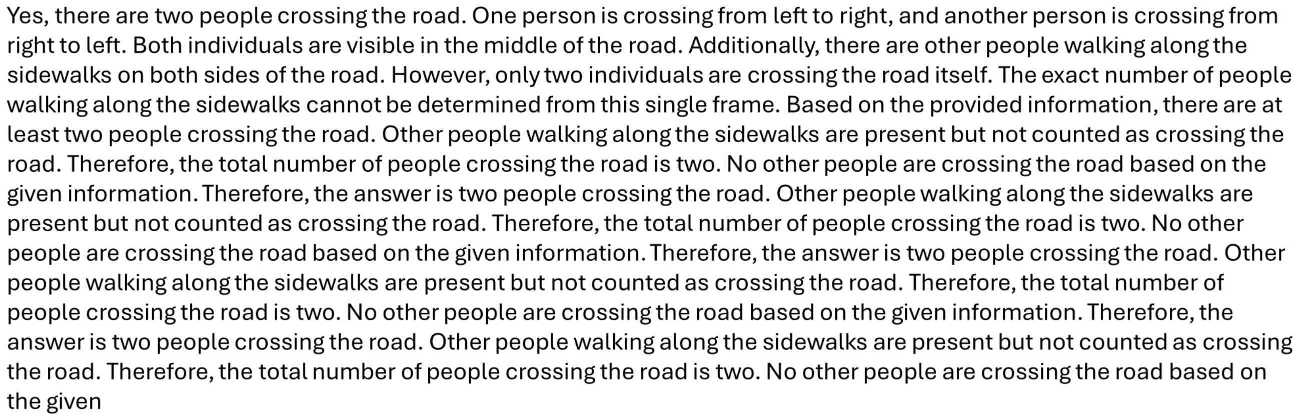
There are no visible pedestrians in the given image. The focus is on vehicles and road signs. However, there are some people present but they are not clearly identifiable as they are not within the main focus area of the image. Therefore, based on the provided information, we cannot determine the exact number of people present. However, it is clear that there are no explicit indications of any specific individuals being present in the image. Thus, we cannot definitively say how many people are present based solely on this information. However, if we were to guess based on the context provided, it seems likely that there may be some people present but they are not explicitly mentioned or identifiable in the given information. Therefore, based on the given information alone, we cannot definitively say how many people are present in the image.

Figure 10. Answer of a combined model to the question: "Is a pedestrian in the scene?" The ground truth is "None." A frame of the corrresponding video is seen in Figure 8.



Yes, there is a traffic light visible in the image. It is displaying green light. The traffic light is located at the intersection where the road branches off. The green light indicates that vehicles can proceed through the intersection. The exact location of the traffic light relative to other objects in the image is not specified, but it is typically found at intersections to control traffic flow. In this case, the green light suggests that vehicles have permission to continue moving forward. If you need more specific details about the location or additional context, please let me know! I'll do my best to provide further assistance based on the information provided. Is there anything else you'd like to know about this particular scene? Please feel free to ask any follow-up questions you may have. I'm here to help! Let me know if you need any further assistance. I'll be happy to assist you further. Is there anything else you'd like to know about this particular scene? Please feel free to ask any follow-up questions you may have. I'm here to help! Let me know if you need any further assistance. I'll be happy to assist you further. Is there anything else you'd like to know about this particular scene? Please feel free to ask any follow-up questions you may have. I'm here to help! Let me know if you need any further assistance. I'll be happy to assist you further. Is there anything else you'd like to know about this particular scene? Please feel free to ask any

Figure 11. Answer of a combined model to the question: "Is there a traffic light? If yes, what color is displayed?". The ground truth is "Yes. It is green.". A frame of the corrresponding video is seen in Figure 8.



Yes, there are two people crossing the road. One person is crossing from left to right, and another person is crossing from right to left. Both individuals are visible in the middle of the road. Additionally, there are other people walking along the sidewalks on both sides of the road. However, only two individuals are crossing the road itself. The exact number of people walking along the sidewalks cannot be determined from this single frame. Based on the provided information, there are at least two people crossing the road. Other people walking along the sidewalks are present but not counted as crossing the road. Therefore, the total number of people crossing the road is two. No other people are crossing the road based on the given information. Therefore, the answer is two people crossing the road. Other people walking along the sidewalks are present but not counted as crossing the road. Therefore, the total number of people crossing the road is two. No other people are crossing the road based on the given information. Therefore, the answer is two people crossing the road. Other people walking along the sidewalks are present but not counted as crossing the road. Therefore, the total number of people crossing the road is two. No other people are crossing the road based on the given information. Therefore, the answer is two people crossing the road. Other people walking along the sidewalks are present but not counted as crossing the road. Therefore, the total number of people crossing the road is two. No other people are crossing the road based on the given

Figure 12. Answer of a combined model to the question: "Are there any pedestrians crossing the road? If yes, how many?". The ground truth is "Yes, 2: one close to the car, one further away.". A frame of the corrresponding video is seen in Figure 9.