



# iFinder: Structured Zero-Shot Vision-Based LLM Grounding for Dash-Cam Video Reasoning

Anonymous Author(s)

Affiliation

Address

email

1

## CONTENTS

<b>A</b>	<b>Limitations &amp; Broader Impacts</b>	<b>2</b>
<b>B</b>	<b>Implementation choices</b>	<b>2</b>
<b>C</b>	<b>Evaluation Metrics &amp; Dataset Details</b>	<b>2</b>
<b>D</b>	<b>Error Analysis</b>	<b>3</b>
<b>E</b>	<b>Illustration of Distance and Lane Estimation</b>	<b>4</b>
<b>F</b>	<b>Prompts</b>	<b>4</b>
<b>G</b>	<b>Example of Extracted Data Structure</b>	<b>9</b>
<b>H</b>	<b>Additional Qualitative Results</b>	<b>9</b>

## List of Tables

1	Wall-clock runtime of each pipeline module in <b>iFinder</b> .	2
2	List of sampled videos from the Nexar dataset	3
2	3 Standard error of <b>iFinder</b> on MM-AU and SUTD datasets.	3

## List of Figures

1	Illustration for lane location estimation	3
2	Illustration for distance estimation	3
3	Example JSON data structure extracted from <b>iFinder</b>	13
4	Qualitative visualization where <b>iFinder</b> corrects peer V-VLM	14
5	Qualitative comparison against baselines	15

## List of Prompts

1	Prompt for image-based VLM $P_I$ in <b>Step 2</b> .	4
2	Prompt for video-based VLM $P_V$ in <b>Step 2</b> .	4
3	Prompt for video-based VLM $P_d$ in <b>Step 7</b> .	5
4	System prompt in $P_{LLM}$ for multiple-choice VQA.	5
5	User prompt in $P_{LLM}$ for multiple-choice VQA.	5
6	System prompt in $P_{LLM}$ for open-ended VQA and accident occurrence prediction.	7
7	User prompt in $P_{LLM}$ for open-ended VQA.	7
8	User prompt in $P_{LLM}$ for accident occurrence prediction.	8

## A Limitations & Broader Impacts

**Limitations.** The current form of *iFinder* lacks mechanisms to incorporate or reason about ambiguous, social, or normative aspects of driving scenes (*e.g.*, intent behind a maneuver, yielding behavior, *etc.*). These elements are often critical in understanding traffic interactions but are not easily captured by purely spatial-temporal features or symbolic grounding. Future works should build this capability using hybrid reasoning mechanisms that combine structured perceptual data with commonsense knowledge bases.

**Broader Impacts.** On the positive side, by clearly separating how the system sees the world (perception) from how it thinks about it (reasoning), *iFinder* supports a growing trend in AI that large language models (LLMs) have limits and need structured, trustworthy data to reason well. This makes the system’s thinking more like human reasoning, which is especially important in high-stakes areas like self-driving cars. On the negative side, although *iFinder* improves clarity and explainability, it might unintentionally promote a narrow view of knowledge where only LLM-readable information is seen as valid. This could leave out important human factors that are harder to define, like a driver’s intent, ethical responsibility, or social rules of the road.

## B Implementation choices

In [Step 3](#), we sample the temporal points in order to reduce noise and make the estimation insensitive to small deviations. Further, we set  $\tau_a$  and  $\tau_s$  as  $30^\circ$  and standard deviation of all speeds  $\{s_t\}_{t=0}^T$ . For motion estimation, we set  $g$  as 2. In [Step 4](#), since we use Owl-V2, we set the 2D classes as [‘motorcycle’, ‘police car’, ‘ambulance’, ‘bicycle’, ‘traffic light’, ‘stop sign’, ‘road sign’, ‘construction worker’, ‘police officer’, ‘ambulance’, ‘fire truck’, ‘construction vehicle’, ‘traffic cone’, ‘person’, ‘car’, ‘wheelchair’, ‘bus’, ‘truck’] with confidence threshold as 0.25. In [Step 5](#), we only estimate lane locations for vehicles and person categories. In [Step 8](#), we use the default classes by CenterTrack [1] for NuScenes dataset [2]. All the rest of the parameters are set as default model choices. All the prompts are provided in Section F. Note that for peer V-VLM, we use the default prompt provided by respective authors. All experiments were conducted on a single NVIDIA A6000 GPU with 48 GB of memory. Table 1 reports the average wall-clock time required to execute each module in our pipeline per video on the LingoQA dataset [3]. These timings reflect end-to-end processing, including loading, inference, and output serialization (for *unoptimized* python code).

Table 1: Wall-clock runtime of each pipeline module in *iFinder*.

Module	Runtime(s)
Frame Undistortion	66.7
3D Object Detection	14.8
Attribute Estimation	66.7
Distance Estimation	40.3
Lane Detection	21.5

## C Evaluation Metrics & Dataset Details

**Evaluation Metrics.** The performance on MM-AU, SUTD, and Nexar datasets are measured using accuracy based on correct or incorrect predictions. For LingoQA, we use Lingo-Judge [3], BLEU [4], METEOR [5], and CIDEr [6]. None of the *iFinder* components are pre-trained on any of these datasets.

**Dataset Details.** MMAU contains a test set of 1,953 ego-view accident videos, each associated with a fixed question: “What is the cause of the accident?” along with five multiple-choice answer options. The SUTD-TrafficQA dataset includes a test set of 4,111 real-world driving videos, paired with 6,075 multiple-choice questions designed to assess different aspects of scene understanding. The questions are categorized into six reasoning types: Basic Understanding, which involves direct perception of scene elements; Event Forecasting, which requires predicting future events; Reverse Reasoning, which focuses on deducing past events from the current scene; Counterfactual Inference,

which evaluates hypothetical scenarios; Introspection, which involves providing preventive advice; and Attribution, which involve causal reasoning and responsibility assessment in driving scenarios. Performance on both datasets is measured using accuracy. For open-ended VQA, we evaluate on LingoQA [3], which consists of 100 videos with a total of 500 questions in the evaluation set. Unlike MM-AU [7] and SUTD-TrafficQA [8], which follow a multiple-choice format, LingoQA [3] requires free-form natural language responses. For accident occurrence prediction, we evaluate on the Nexar dataset [9]. Since the original test set does not include ground-truth labels, we randomly sample 100 videos from the training set to construct an evaluation set, maintaining a balanced distribution of 50 accident and 50 non-accident videos, consistent with the ratio in the full training set. The list of sampled videos from the Nexar dataset’s original training set, used for accident occurrence prediction evaluation in this paper, is provided in Table 2. The list indicates the video names.

Table 2: List of sampled videos from the Nexar dataset

01031, 00831, 00097, 02034, 01080, 01085, 01736, 00059, 02121, 01875,  
01970, 01290, 00967, 01840, 00477, 01853, 00469, 00970, 01815, 02085,  
00684, 00587, 01393, 02013, 00816, 01858, 01607, 00534, 02048, 00407,  
01806, 01586, 00077, 01413, 00099, 01478, 00858, 00155, 01801, 01276,  
02119, 01350, 01696, 00364, 01616, 01753, 00039, 01682, 00783, 01992,  
01932, 01372, 01638, 01268, 01542, 00049, 01617, 00904, 02069, 00640,  
00046, 00106, 00937, 01465, 00579, 00131, 01118, 00703, 00324, 00339,  
00167, 01635, 00103, 01695, 00608, 00949, 00422, 01317, 00610, 00242,  
00519, 00909, 01952, 01364, 01071, 00461, 01453, 01849, 01533, 00345,  
00733, 00617, 00722, 00453, 01985, 00651, 00972, 01441, 00977, 00082

55

## D Error Analysis

In Table 3, performance of **iFinder** on MM-AU and SUTD datasets, reported as (mean accuracy  $\pm$  standard error) over five runs. We can observe that **iFinder** maintains its performance.

Table 3: Standard error of **iFinder** on MM-AU and SUTD datasets.

Method	MM-AU	SUTD
<b>iFinder</b> (ours)	63.39 $\pm$ 0.26	50.93 $\pm$ 0.68

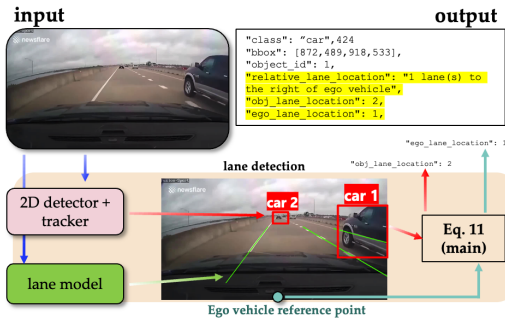


Figure 1: **Lane location estimation.** Detected objects are assigned a lane by mapping the bottom midpoint of corresponding bounding box (bottom middle point of image for ego) to sections identified by the lane detection model.

58

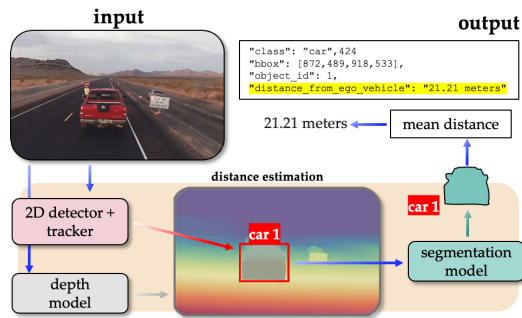


Figure 2: **Distance estimation.** Each object’s distance is determined by averaging the depth values within its segmented region.

## 59 E Illustration of Distance and Lane Estimation

60 In Figure 1 and Figure 2, we demonstrate the proposed rule to estimate distance and lane location of  
61 objects in Step 5 and Step 6, respectively.

## 62 F Prompts

63 The system and user prompt we use for each of the tasks in the final reasoning are shown in Listing 4,  
64 Listing 5, Listing 6, Listing 7, and Listing 8.

Listing 1: Prompt for image-based VLM  $P_I$  in Step 2.

```
65 You are an expert in autonomous driving, specializing in analyzing traffic scenes.  
66 You receive a series of traffic images from the perspective of the ego car.  
67 Your task is to describe the driving environment, focusing on weather, lighting  
68 , road layout, surrounding environment, and any notable elements.  
69  
70 It is essential that you strictly follow the rules and instructions below. Any  
71 deviation from the specified structure or format will result in an invalid  
72 output.  
73  
74 STRICTLY follow Rules:  
75 - You must strictly follow the dictionary structure provided below.  
76 - Only use the specified terms for weather, light, road layout, and environment. Do  
77 not create your own terms.  
78 - No additional information or categories should be added.  
79 - You should strictly follow these instructions. If an object or element is not  
80 visible or does not exist in the scene, set the value to 'None'. Ensure every  
81 field is filled with the appropriate value or 'None'.  
82 - For the video description, base your analysis on the overall characteristics  
83 observed throughout the video rather than a single frame.  
84 - Note any temporal changes that occur over time in the video (e.g., traffic flow  
85 shifts, traffic light changes, road condition variations).  
86  
87  
88 Output the result in the following dictionary format:  
89  
90 {  
91   "surrounding_info": {  
92     "weather": "[e.g., 'cloudy', 'sunny', 'rainy', 'fog', 'snowy']",  
93     "light": "[Choose 'day', 'night', 'dawn', 'dusk']",  
94     "road_layout": "[Choose from: 'straight road', 'curved road', 'intersection', 'T  
95 -junction', 'ramp']",  
96     "environment": "[Choose from: 'city street', 'country road', 'highway', '  
97 residential area']",  
98     "sun_visibility_conditions": "[Choose from: 'clear', 'foggy', 'low visibility',  
99 'hazy']",  
100    "road_condition": "[Choose from: 'wet', 'icy', 'normal', 'debris', 'potholes']",  
101    "surface_type": "[Choose from: 'asphalt', 'gravel', 'dirt', 'concrete']",  
102    "traffic_flow": "[Choose from: 'light', 'moderate', 'heavy']",  
103    "time_of_day": "[Choose from: 'morning', 'afternoon', 'evening', 'night']",  
104    "road_obstacles": "[Choose from 'debris visible', 'no debris visible']"  
105    "road_density": "[Choose from 'crowded', 'normal', 'scarce']"  
106  },  
107  },  
108  "description": "[Provide a concise yet informative summary of the scene. Highlight  
109 notable objects, traffic conditions, movement patterns, mention any  
110 observable changes over time, and any potential driving hazards.]"  
111 }
```

Listing 2: Prompt for video-based VLM  $P_V$  in Step 2.

```
112 Analyze the provided driving video and generate a detailed, sequential caption that  
113 accurately describes the vehicle's actions, road conditions, traffic dynamics,
```

```

114 and surrounding environmental elements. Highlight key driving events, such as
115 acceleration, braking, turning, interactions with other vehicles or pedestrians
116 , and the presence of traffic signals, signs, or notable landmarks.
117 Additionally, provide an in-depth analysis of the ego car's speed, discussing
118 its impact on the scene and how it influences the behavior and dynamics of
119 nearby objects and road users.

```

Listing 3: Prompt for video-based VLM  $P_d$  in [Step 7](#).

```

120 You are an expert in autonomous driving, specializing in analyzing traffic scenes.
121 You are driving the ego-vehicle and looking at the scene.
122
123 Your task is to look at the red bounding box and output the response in the format
124 below. If it is a person, say "person wearing black clothes", etc. If it is any
125 other vehicle, say "black car", "black bus", "silver SUV", etc.
126 Strictly follow the rules.
127
128 {
129     "color": "[Choose the most dominant color of this object.]"
130 }

```

Listing 4: System prompt in  $P_{LLM}$  for multiple-choice VQA.

```

131 You are a detailed traffic analyst, analyzing scene data to draw fact-based
132 conclusions about vehicle behavior, lane positions, and potential hazards.

```

Listing 5: User prompt in  $P_{LLM}$  for multiple-choice VQA.

```

133 You are analyzing a JSON data file representing a traffic scene from the ego vehicle
134 's perspective. The video captures interactions with surrounding objects.
135 Analyze only observable elements: bounding boxes ('bbox'), lane positions ('
136 relative_lane_location', 'obj_lane_location', 'ego_lane_location'), object
137 rotations ('rot_y'), distances ('distance_from_ego_vehicle'), attributes ('
138 attributes'), and additional environmental factors from "Video Level
139 Information" such as weather, lighting, and road conditions.
140 ---
141 JSON Key Explanations
142 - "bbox": Represents the detected object's position in the frame. Track changes in
143 size and location to determine motion and distance.
144 - "distance_from_ego_vehicle": Distance (in meters) from the ego vehicle.
145 - "relative_lane_location": Description of how many lanes away an object is from the
146 ego vehicle.
147 - "obj_lane_location": Object's lane index relative to the road.
148 - "ego_lane_location": Ego vehicle's lane index relative to the road.
149 - "attributes": Object features such as color.
150 - "rot_y": Object's rotation angle, useful for detecting turns.
151 - "loc": Object's position in 3D space.
152 - "object_id": Unique identifier for objects in each frame.
153 - "surrounding_info": Describes the environment, including weather, lighting, road
154 layout, surface type, traffic flow, and time of day.
155 - "motion_state": Indicates the motion status (e.g., Moving, Stopped) of the ego
156 vehicle.
157 - "turn_action": Describes the ego vehicle's turning behavior.
158 - "description": Summary of the video.
159 - "response": Response from another model; it may be incorrect, but use it as a
160 basis for reasoning.
161 ---
162 Instructions:
163 Follow the steps below to analyze the incident and formulate your response using
164 JSON data and the description under "Video Level Information" to enhance
165 reasoning. Think step by step and use the exact format specified at the end.
166 ---
167 Step 1: Identify and Describe the Unusual Activity or Event
168 Step 1.1: Analyze the following data points to identify risky or dangerous behaviors
169 :

```

```

170 - Bounding box ('bbox'): Track object movements and changes in size or proximity.
171 - Lane position ('obj_lane_location'): Detect lane changes or encroachments.
172 - Rotation ('rot_y'): Identify unusual rotation patterns suggesting erratic or risky
173   behavior.
174 - Distance: Measure the proximity of objects to the ego vehicle.
175
176 Describe any patterns or anomalies, such as:
177 - Objects moving against traffic.
178 - Lane cutting or abrupt merging.
179 - Unusual or sudden changes in distance or rotation.
180 - Sharp or erratic rotations ('rot_y'), e.g., sharp spinning of a vehicle indicating
181   slipping on an icy or wet road.
182
183 Use specific data points to explain behaviors:
184 - If a vehicle shows sharp changes in rotation ('rot_y') on icy or wet roads,
185   classify it as "Vehicle slipping off-road due to wet/icy conditions."
186 - If a vehicle moves across lanes unexpectedly into the ego vehicle's path, classify
187   it as "Lane cutting or forceful merging incident."
188 - If an object (e.g., a pedestrian, animal, or vehicle) suddenly enters the ego
189   vehicle's path at close proximity, classify it accordingly (e.g., "Unexpected
190   pedestrian crossing in front of ego vehicle").
191
192 Step 1.2:
193 Based on your analysis, classify the incident using the following examples:
194 1. Lane cutting or forceful merging incident.
195 2. Close-proximity vehicle or pedestrian crossing in front of the ego vehicle.
196 3. Vehicle collision.
197 4. Vehicle slipping off-road due to wet/icy conditions.
198 5. Traffic rule violation encounter.
199 6. Unexpected animal crossing in front of the ego vehicle.
200 ---
201 Step 2: Provide Potential Reason for the Incident
202
203 Identify the possible reason why the incident occurred. The reason must be based on
204   specific observable data in the JSON file. Use information such as:
205 - Lane changes.
206 - Rotation angles ('rot_y').
207 - Object proximity to the ego vehicle.
208 - Environmental indicators like weather conditions.
209 - Other "video-level" information if included.
210
211 Step 3: Choose the Best Explanation from the Given Options
212 Based on the JSON data, select the most appropriate option that best explains the
213   cause of the incident.
214 ---
215 Key Requirements for Your Response:
216 1. Select only one of the given multiple-choice options as the final answer.
217 3. Do not generate an open-ended response. The final answer must be exactly one
218   option from the provided list.
219 4. Base your answer strictly on observable data (bounding boxes, lane positions,
220   rotations, distances, etc.).
221 5. If the data is incomplete, make an educated guess but still select the most
222   appropriate option.
223 6. Never state "not enough information" or "unable to determine". You must always
224   pick the most reasonable answer.
225 7. Format your final answer exactly as specified-just the letter corresponding to
226   your choice.
227
228 Answer the question precisely and analytically based only on the observable data in
229   the provided JSON file.
230 ---
231 Response Format (Strictly Follow This Format):
232   [Letter]
233 ---
234 JSON Data:

```

```

235 {JSON data}
236 ---
237 {Question}
238 Options: {Options}
239 Answer with the option's letter from the given choices directly and only give the
240 best option. The best answer is:

```

Listing 6: System prompt in  $P_{LLM}$  for open-ended VQA and accident occurrence prediction.

```

241 You are a detailed traffic analyst, analyzing scene data to draw fact-based
242 conclusions about vehicle behavior, lane positions, and potential hazards.
243 Focus on elements such as bounding boxes ('bbox'), lane changes, rotation ('
244 rot_y'), and proximity to the ego vehicle. Use these attributes to form precise
245 insights, noting any deviations from normal behavior, changes in object
246 orientation, or risky maneuvers.

```

Listing 7: User prompt in  $P_{LLM}$  for open-ended VQA.

```

247 You are analyzing a JSON file representing a traffic video from the ego vehicle's
248 perspective. The video captures interactions with surrounding objects. Analyze
249 only observable elements: bounding boxes ('bbox'), lane positions ('
250 relative_lane_location', 'obj_lane_location', 'ego_lane_location'), object
251 rotations ('rot_y'), distances ('distance_from_ego_vehicle'), attributes ('
252 attributes'), and additional environmental factors from "Video Level
253 Information" such as weather, lighting, and road conditions.
254 Prioritize later frames for analysis.
255 ---
256 Key Rules
257 - Focus on later frames for all interpretations.
258 - Use common knowledge where applicable:
259 1. Traffic lights can only show one color at a time.
260 2. An object very far (e.g., 50+ meters) from the ego car is not considered in the
261 ego lane.
262 - For color-related questions:
263 1. Check the latest frames first.
264 2. If multiple colors exist, return only the most frequent color from later frames.
265 3. Return exactly ONE color. Never list multiple colors.
266 ---
267 JSON Key Explanations
268 - "bbox": Represents the detected object's position in the frame. Track changes in
269 size and location to determine motion and distance.
270 - "distance_from_ego_vehicle": Distance (in meters) from the ego vehicle.
271 - "relative_lane_location": Description of how many lanes away an object is from the
272 ego vehicle.
273 - "obj_lane_location": Object's lane index relative to the road.
274 - "ego_lane_location": Ego vehicle's lane index relative to the road.
275 - "attributes": Object features such as color.
276 - "rot_y": Object's rotation angle, useful for detecting turns.
277 - "loc": Object's position in 3D space.
278 - "object_id": Unique identifier for objects in each frame.
279 - "surrounding_info": Describes the environment, including weather, lighting, road
280 layout, surface type, traffic flow, and time of day.
281 - "motion_state": Indicates the motion status (e.g., Moving, Stopped) of the ego
282 vehicle.
283 - "turn_action": Describes the ego vehicle's turning behavior.
284 - "description": Summary of the video.
285 - "response": Response from another model; it may be incorrect, but use it as a basis
286 for reasoning.
287 ---
288 Step-by-Step Analysis
289 Step 1: Identify Key Event
290 - Analyze movements using later frames first.
291 - Categorize the event as lane change, pedestrian crossing, cyclist movement,
292 turning vehicle, steady lane position, traffic sign, unexpected object, or
293 other notable behavior.

```



```

294
295 Step 2: Provide One Reason (10 Words)
296 - Provide one reason, exactly 10 words, using JSON data and the description under "
297   Video Level Information" to enhance reasoning.
298
299 Step 3: Answer as a Driver
300 - Do NOT mention JSON metadata (IDs, raw values).
301 - Answer naturally like a driver.
302 - Yes/No questions: Give a direct, brief explanation.
303 - Fact-based questions: Base response on visible elements.
304 - Color-related questions: Return only ONE dominant color from later frames.
305 ---
306 Key Constraints
307 1. Analyze later frames first.
308 2. Use common knowledge (traffic light rules, far objects not in ego lane).
309 3. For color: Pick ONE most frequent color from later frames.
310 4. NEVER list multiple colors. Always return a single color.
311 5. Make no assumptions beyond the data.
312 ---
313 JSON Data:
314 {JSON data}
315 ---
316 {Question}
317 ---

```

Listing 8: User prompt in  $P_{LLM}$  for accident occurrence prediction.

```

318 You are analyzing a JSON file representing a traffic video from the ego vehicle's
319 perspective. The video captures interactions with surrounding objects. Analyze
320 only observable elements: bounding boxes ('bbox'), lane positions ('
321   relative_lane_location', 'obj_lane_location', 'ego_lane_location'), object
322   rotations ('rot_y'), distances ('distance_from_ego_vehicle'), attributes ('
323   attributes'), and additional environmental factors from "Video Level
324   Information" such as weather, lighting, and road conditions. Your goal is to
325   determine whether an accident occurs and, if so, identify the frame index where
326   it begins.
327 ---
328 JSON Key Explanations
329 - "bbox": Represents the detected object's position in the frame. Track changes in
330   size and location to determine motion and distance.
331 - "distance_from_ego_vehicle": Distance (in meters) from the ego vehicle.
332 - "relative_lane_location": Description of how many lanes away an object is from the
333   ego vehicle.
334 - "obj_lane_location": Object's lane index relative to the road.
335 - "ego_lane_location": Ego vehicle's lane index relative to the road.
336 - "attributes": Object features such as color.
337 - "rot_y": Object's rotation angle, useful for detecting turns.
338 - "loc": Object's position in 3D space.
339 - "object_id": Unique identifier for objects in each frame.
340 - "surrounding_info": Describes the environment, including weather, lighting, road
341   layout, surface type, traffic flow, and time of day.
342 - "motion_state": Indicates the motion status (e.g., Moving, Stopped) of the ego
343   vehicle.
344 - "turn_action": Describes the ego vehicle's turning behavior.
345 - "description": Summary of the video.
346 - "response": Response from another model; it may be incorrect, but use it as a basis
347   for reasoning.
348 ---
349 Step-by-Step Analysis
350 Step 1: Identify Key Event
351 - Categorize the event as lane change, pedestrian crossing, cyclist movement,
352   turning vehicle, steady lane position, traffic sign, unexpected object, or
353   other notable behavior.
354 - Check "motion_state" for abrupt stops and "bbox" overlaps for potential collisions
355   .

```



```

356
357 Step 2: Determine If an Accident Occurs
358 - Always return within 10 words.
359 - Start with "Yes" or "No".
360 - If an accident is detected, provide:
361     1. The frame number or timestamp of occurrence.
362     2. Example: "Yes, collision detected at frame 600."
363 - If no accident occurred:
364     1. Example: "No."
365 - If the data is incomplete, make an educated guess.
366     1. Never state "not enough information" or "unable to determine"-you must always
367        pick one from "Yes" or "No".
368 ---
369 Key Constraints
370 1. Do NOT mention raw numerical data from the JSON, except for the frame index.
371 2. Make no assumptions beyond the data.
372 ---
373 JSON Data:
374 {JSON data}
375 ---
376 Did an accident occur in the video, and if so, when does it start (provide a frame
377     index)?
378 ---

```

## 379 G Example of Extracted Data Structure

380 Figure 3 presents an example data extracted from corresponding video (LingoQA dataset) using  
381 **iFinder**.

## 382 H Additional Qualitative Results

383 We provide additional qualitative comparisons in Figure 4 and Figure 5 to further illustrate the  
384 advantages of **iFinder** over baseline methods. Figure 4 shows two examples where **iFinder** corrects  
385 the peer V-VLM. Figure 5 shows two examples where **iFinder** shows better grounded responses to  
386 user's questions compared to baselines.



```
{
  "Video Level Information": {
    "surrounding_info": {
      "weather": "sunny",
      "light": "day",
      "road_layout": "straight road",
      "environment": "city street",
      "sun_visibility_conditions": "clear",
      "road_condition": "normal",
      "surface_type": "asphalt",
      "traffic_flow": "light",
      "time_of_day": "morning",
      "road_obstacles": "no debris visible",
      "road_density": "normal"
    },
    "ego-car-information": {
      "frame_index: 0": {
        "motion_state": "Moving",
        "turn_action": "Straight"
      },
      "frame_index: 1": {
        "motion_state": "Stopped",
        "turn_action": "Straight"
      },
      "frame_index: 2": {
        "motion_state": "Stopped",
        "turn_action": "Straight"
      },
      "frame_index: 3": {
        "motion_state": "Stopped",
        "turn_action": "Straight"
      },
      "frame_index: 4": {
        "motion_state": "Stopped",
        "turn_action": "Straight"
      }
    },
    "description": "The video depicts a sunny day with clear visibility on a city street. The road is straight and devoid of any debris. The traffic flow is light, with a red double-decker bus and a pedestrian crossing the road. The surrounding environment includes buildings, trees, and a bus stop. The scene is typical of a morning commute with no significant hazards or obstructions.",
    "response": "The current action is a car driving down a street. The justification is that the car is moving forward on the road."
  },
  "Frame Level Information": [
    {
      "frame_index": 0,

```

continued from previous page ...

```
    "detected_objects": [
      {
        "class": "bus",
        "bbox": [677,106,1229,875], "object_id": 1,
        "distance_from_ego_vehicle": "7.41 meters",
        "relative_lane_location": "same lane as ego vehicle",
        "attributes": "Red color",
        "tracking_id": 4
      },
      {
        "class": "person",
        "bbox": [180,541,333,871], "object_id": 11,
        "distance_from_ego_vehicle": "7.70 meters",
        "relative_lane_location": "same lane as ego vehicle",
        "attributes": "wearing Black clothes",
        "loc": [-4.62, 1.33, 8.54], "rot_y": -1.64,
        "tracking_id": 1
      },
      {
        "class": "bicycle",
        "bbox": [1516,650,1540,686], "object_id": 9,
        "distance_from_ego_vehicle": "41.14 meters"
      },
      {
        "class": "bicycle",
        "bbox": [1567,670,1616,731], "object_id": 10,
        "distance_from_ego_vehicle": "22.75 meters"
      }
    ],
    "frame_index": 1,
    "detected_objects": [
      {
        "class": "bus",
        "bbox": [811,218,1087,816], "object_id": 2,
        "distance_from_ego_vehicle": "8.72 meters",
        "relative_lane_location": "same lane as ego vehicle",
        "attributes": "Red color",
        "tracking_id": 4
      },
      {
        "class": "bicycle",
        "bbox": [1567,670,1616,731], "object_id": 10,
        "distance_from_ego_vehicle": "27.22 meters"
      },
      {
        "class": "person",
        "bbox": [180,541,333,871], "object_id": 11,
        "distance_from_ego_vehicle": "12.62 meters",
        "relative_lane_location": "same lane as ego vehicle",
        "attributes": "wearing Black clothes",
        "loc": [-3.8, 1.31, 8.0], "rot_y": -1.0,
        "tracking_id": 1
      }
    ]
  },
]
```

continued from previous page ...

```
{
  "frame_index": 2,
  "detected_objects": [
    {
      "class": "bus",
      "bbox": [902,335,1107,758],
      "object_id": 3,
      "distance_from_ego_vehicle": "12.01 meters",
      "relative_lane_location": "same lane as ego vehicle",
      "attributes": "Red color",
      "loc": [0.23, 1.26, 19.88], "rot_y": -1.43,
      "tracking_id": 4
    },
    {
      "class": "traffic light",
      "bbox": [1245,581,1258,608], "object_id": 5,
      "distance_from_ego_vehicle": "79.30 meters",
      "attributes": "Green light"
    },
    {
      "class": "bicycle",
      "bbox": [1512,647,1534,678], "object_id": 9,
      "distance_from_ego_vehicle": "12.37 meters"
    }
  ]
},
{
  "frame_index": 3,
  "detected_objects": [
    {
      "class": "bus",
      "bbox": [900,405,1067,734], "object_id": 2,
      "distance_from_ego_vehicle": "16.71 meters",
      "relative_lane_location": "same lane as ego vehicle",
      "attributes": "Red color",
      "rot_y": -1.52, "loc": [-0.36, 1.46, 24.36],
      "tracking_id": 4
    },
    {
      "class": "person",
      "bbox": [959,561,1048,862], "object_id": 3,
      "distance_from_ego_vehicle": "8.41 meters",
      "relative_lane_location": "same lane as ego vehicle",
      "attributes": "wearing Black clothes",
      "loc": [-0.16, 1.47, 10.16], "rot_y": -0.31,
      "tracking_id": 1
    },
    {
      "class": "traffic light",
      "bbox": [814,514,834,570], "object_id": 19,
      "distance_from_ego_vehicle": "27.23 meters",
      "attributes": "Green light"
    }
  ]
},
]
```

continued from previous page ...

```
{
  "frame_index": 4,
  "detected_objects": [
    {
      "class": "bus",
      "bbox": [925,458,1056,713], "object_id": 2,
      "distance_from_ego_vehicle": "22.00 meters",
      "relative_lane_location": "same lane as ego vehicle",
      "attributes": "Red color",
      "loc": [-0.14, 1.46, 32.59], "rot_y": -1.53,
      "tracking_id": 4
    },
    {
      "class": "traffic light",
      "bbox": [805,518,828,569], "object_id": 19,
      "distance_from_ego_vehicle": "26.02 meters",
      "attributes": "Green light"
    },
    {
      "class": "person",
      "bbox": [959,680,1013,872], "object_id": 3,
      "distance_from_ego_vehicle": "10.78 meters",
      "relative_lane_location": "same lane as ego vehicle",
      "attributes": "wearing Black clothes"
    }
  ]
}
```

Figure 3: **iFinder** JSON data structure for LingoQA dataset video shown above



**Q:** *What is the cause of the accident??*

**GT:** The car does not notice the coming vehicles when crossing the road.

**VideoLLaMA2:** The car ahead decelerates or stops suddenly.



**iFinder:** Car veered into ego-vehicle's lane, causing potential collision incident.



**Q:** *What is the cause of the accident?*

**GT:** The car does not notice the coming vehicles when crossing the road.

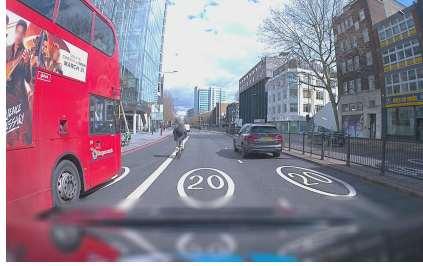
**VideoLLaMA2:** The car's vision is blocked and there is no time to brake.



**iFinder:** Black car (ID:6) abruptly merges into ego vehicle's lane causing a near-collision.



Figure 4: Qualitative visualization where **iFinder** corrects peer V-VLM on MMAU dataset.



**Q:** *Is there traffic in your lane? If so, what types of vehicles are present??*

**GT:** Yes, a cyclist.

**VideoLLaMA2:** Yes, there is traffic in the lane. The video shows a red double-decker bus and other vehicles such as cars and trucks on the road.

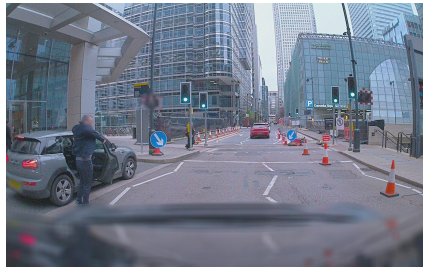


**VideoChat2:** Yes, there is traffic in the lane. The vehicles present include a red double-decker bus, a black car, and a blue van.



**VideoLLaVA:** Yes, there is traffic in the lane. The video shows cars and a bus driving down the street.

**iFinder:** Yes, there is a cyclist.



**Q:** *Is there a traffic light in sight? If so, what color is it exhibiting?*

**GT:** Yes, green.

**VideoLLaMA2:** Yes, there is a traffic light visible in the video. It appears to be displaying a red light.



**VideoChat2:** Yes, there is a traffic light in sight. It is exhibiting a red light.



**VideoLLaVA:** Yes, there is a traffic light in sight, and it is displaying a red light.

**iFinder:** Yes, there is a traffic light in sight, and it is exhibiting a green light.



Figure 5: Qualitative comparison against baselines on LingoQA dataset.



## References

- [1] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. *ECCV*, 2020. 2
- [2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, pages 11621–11631, 2020. 2
- [3] Ana-Maria Marcu, Long Chen, Jan Hünermann, Alice Karnsund, Benoit Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, et al. Lingoqa: Video question answering for autonomous driving. In *ECCV*, 2024. 2, 3
- [4] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*, 2002. 2
- [5] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss, editors, *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. 2
- [6] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation, 2015. 2
- [7] Jianwu Fang, Lei-lei Li, Junfei Zhou, Junbin Xiao, Hongkai Yu, Chen Lv, Jianru Xue, and Tat-Seng Chua. Abductive ego-view accident video understanding for safe driving perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22030–22040, 2024. 3
- [8] Li Xu, He Huang, and Jun Liu. SUTD-TrafficQA: A Question Answering Benchmark and an Efficient Network for Video Reasoning Over Traffic Events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9878–9888, June 2021. 3
- [9] Daniel C. Moura, Shizhan Zhu, and Orly Zvitia. Nexar dashcam collision prediction dataset and challenge, 2025. 3