1. VQA Generation Pipeline

1.1. Scenario Aggregation from Multiple Sources

Scene collection of nuScenes dataset. To collect nuScenes scenarios with the original observations (nuScenes-real), we use the Python implementation of **nuscenes-devkit** [1] to explore traffic scenarios. Following the naming paradigm provided in the official nuScenes documentation, for each `sample` in a `scene`, we extract its `CAM_FRONT sampled_data`. At this point, we can associate a recorded keyframe with its image observation. In the first filtering process, if an object is annotated with than a "3" level of visibility or is scanned by less than five rays of Lidar, then it is considered "invisible," and its information will not be recorded. However, this first pass doesn't consider visual occlusion since the visibility of objects is annotated on the scene level instead of the frame level in nuScenes. Therefore, a second filtering pass is instigated. The **nuscenes-devkit** provides API to project 3D bounding boxes(8 vertices) of objects onto the 2D image observations, and we create the maximum enclosing 2D filled bounding boxes of the 8 vertices. Then, these filled rectangles are painted onto a black image following the distance order of objects to mimic the process of z-buffering, and boxes of distant objects will be overlayed by closer objects. Finally, we filter out objects with less than 50% of their 2D boxes visible in the composed image, completing the second filtering pass. The third and last filtering pass removes miscellaneous objects such as `debris` and `vegetation` from objects of interest. After these three filtering stages, the interested objects set will have the information mentioned in the color box to the right recorded. Notably, the nuScenes dataset doesn't have the "color" annotation, and we leave this field empty while collecting the scenarios.

Scene reconstruction with simulator. Leveraging the MetaDrive [2] simulator and ScenarioNet [3] data platforms, we aggregate nuScenes [1] and Waymo [4]. For simulator-reconstructed traffic scenarios, we record frames every five steps (0.5 seconds wall time) until the end. We set a camera with a 60-degree field-of-view and 1920 $\times$ 1080 resolution to extract rendering. At each simulation step, we record the following information about the ego and objects within 75 meters of the ego:

`Information Recorded per Frame:`
id, assigned by the simulator.
color, bound to the 3D asset.
height, bound to the 3D asset.
type, bound to the 3D asset.
bounding box in world coordinates.
heading vector in world coordinates.
speed of the object in meters per second (m/s).
position of the center point in world coordinates.
ego camera that observes the vehicle (if any).
visibility of the object to the ego vehicle.
collided objects (if any) at this moment.

Note that if an object is "visible," the camera must capture at least 1,200 pixels of its body. This is implemented by assigning an ID color to each active object in the simulation, and we use a special instance segmentation camera (the same intrinsic and placement as the capturing camera) to capture the ID-color-based rendering. The traffic collected has the following statistics.

Constuction of 3D scene graphs. Each scene graph comprises nodes connected by directed edges representing relative spatial relationships. Each node corresponds to a visible object from the frame information recorded from the previous step, and intrinsic properties (*e.g.*color, height) are contained in the node. Given a reference vector `V`, we determine the relative spatial relationships between current node `A` and node `B` by:

`Relative Spatial Relationships(box A,B;front vector V):`

left or right. Refer to Fig. 1, and we determine the leftmost and rightmost vertices of bounding box A using the reference vector V as the front direction. Then, if all vertices of bounding box B are to the left of the leftmost vertex of A, then we consider B's sidedness to be "left" (and similarly for sidedness to be "right"). If bounding box B satisfies neither of the two conditions, then we consider B's "sidedness" to be "none".

front or back. We determine this relationship similarly to determining "left" or "right", with the modification that V is the left direction.

This reference vector V is the heading of the ego vehicle when determining "left or right", and it's rotated 90 degrees counterclockwise with respect to the yaw axis when determining "front or back". Once we have the two values for

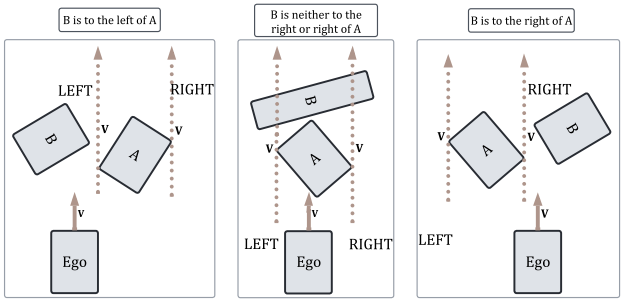


Figure 1. **Top-down illustration of sidedness.** We demand all vertices of box B reside in the "LEFT" region of A for B to be considered "to the left of" (and similarly for "to the right of") A.

"left or right" and "front or back", we draw the corresponding directed edge from A to B from the following:

```
Named Spatial Edges:
l, corresponding to "to the left of."
lb, corresponding to "to the left and behind."
lf, corresponding to "to the left and in front of."
b, corresponding to "behind."
f, corresponding to "in front of."
r, corresponding to "to the right."
rb, corresponding to "to the right and behind."
rf, corresponding to "to the right and in front of."
```

For example, if "l" edge is chosen, this means "B is to the left of A."

1.2. Set-of-Mark Prompting

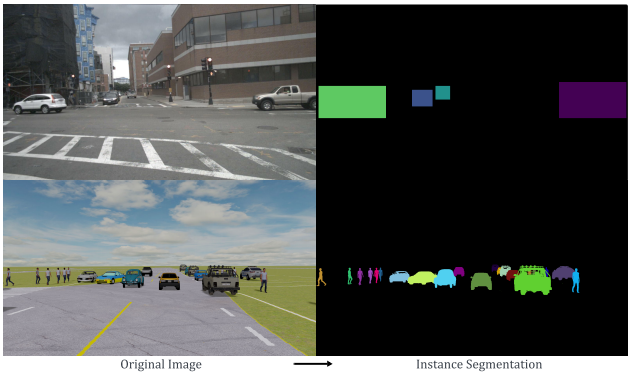


Figure 2. **Instance segmentation masks.** Approximated instance segmentation is generated for real images from the nuScenes dataset. Simulated images are paired with precise instance segmentation.

From Sec. 1.1, we have collected image observations and the corresponding instance segmentation in approximated boxes(nuScenes images) or shape-precise masks(simulated images), as shown in Fig. 2. Then, we run the algorithm illustrated in Fig. 3 adopted from the original Set-of-Mark paper [5] to determine the appropriate position for object labels:

```
// Find center for a region
def Find_Center(r)
    D = DT(r) // Run distance transform
    c = arg max(D) // Find maxium location
    return c
// The main function
def Mark_Allocation(R):
    R̂ = Sorted(R) // Sort regions in ascending
        order of areas
    for k in range(K): do
        r_k = R̂[k] & ¬R̂[: k − 1].sum(0) // Exclude
            k − 1 regions
        C[k] = Find_Center(r_k)
    end
    return C
```

Figure 3. **Labeling algorithm adopted from the Set-of-Mark paper.** Credit to the Set-of-Mark authors.

The Set-of-Mark paper suggested various schemes to perform the visual prompting. For example, using instance segmentation masks and contours are both valid schemes to improve the visual grounding capabilities of vision-language models (VLMs). As mentioned in the main paper, we conducted an ablation study on different prompting schemes to determine the optimal scheme for referral clarity using labels. Using Qwen2 [6] as the zero-shot evaluating model, we fix the prompting scheme with bounding-box annotations, black text background color, and a text size of 1.00 (to reduce label occlusions). The bounding boxes and texts use colors identical to that of the instance segmentation masks of corresponding objects. We use `cv2.rectangle` to draw the bounding boxes onto original images with `thickness = 2`, and we use `cv2.putText` with `font_size = 1` and `thickness = 2`. In addition, we slightly relocate the labels if their corresponding 2D bounding boxes enclose regions less than 1,600 pixels. This is to ensure the visibility of highlighted objects after the visual prompting. The concrete code implementations will be released.

1.3. Question-Answer Generation

1.3.1. Question Generation

MetaVQA adopts a template-based question generation process. Each type of question is bonded to a single template with varying numbers and types of parameters to be replaced with concrete values. We categorize questions into "non-parameterized" and "parameterized " based on the number of parameter types in the template.

Parameterized question generation. Parameters are present for the templates of these questions. These parameters will be replaced upon question generation with concrete values selected from corresponding parameter spaces, the summary of which is provided in Fig. 4. The generation process for a parameterized question is illustrated in Fig. 5: the template of `identify_distance` contains a single `<id1>` parameter, the parameter space of which is all valid labels generated from the Set-of-Mark prompting. In this example, `<id1>` is replaced by the randomly selected label `<0>`. Additionally, multiple parameters belonging to different types can co-exist in a single-question template. Refers to Fig. 6 for an illustration. Observe how concrete values for parameters are sampled from the parameter spaces.

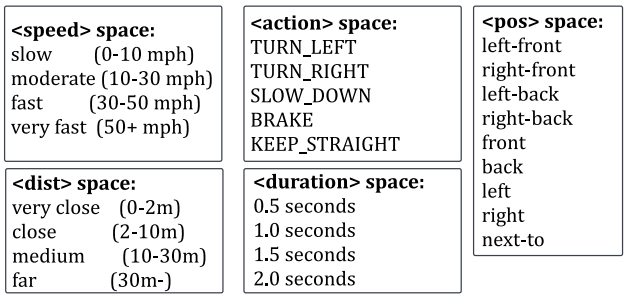


Figure 4. **Parameter space summary**. Note that space of `<id*>` is scenario-dependent, namely, all valid labels.

Non-parameterized question generation. These questions don't have any parameters in their templates, as they demand the VLMs to examine all present objects in observations before answering. An example can be found in Fig. 7. Therefore, no computation is done in the question generation phase.

1.3.2. Answer Generation

A unique query program is selected to generate answers for each type of question. Refer to Fig. 5, Fig. 6, and Fig. 7 for examples. Upon the execution of these query programs, the concrete answers are extracted utilizing scenario information for simulated dynamics. Note that both the question-answer pairs at this stage are not formulated in the multiple-choice setting, and the next stage will reformat the pairs.

1.3.3. Post-processing

At this point, question-answer pairs are already generated. The remaining works are (1) the generation of non-answer candidates for multiple-choice setup (2) the creation of the multiple-choice description strings which map choices with concrete answer candidates, (3) the creation of optional "explanation" strings to elevate VLMs' learning. Each question has different search spaces for non-answer candidates. As shown in Fig. 5, `identify_distance`'s candidate space is the `<dist>` space listed in Fig. 4, while that of `embodied_sideness` is a subset of `<pos>` space, shown in Fig. 6. When applicable, non-answer candidates are selected to challenge the evaluated VLMs maximally. For example, candidate generation in question `identify_type` prioritizes ones present in the scenarios on which the question is constructed. After the candidates' generation, they are put into multiple-choice format as suffixes to the original question, and the answer is replaced by the answer choice. The optional "explanation" strings (used interchangeably with "reasoning") are also programmatically created, depending on the choice-candidate mapping. Complete implementation will be included in the released codebase.

2. MetaVQA Dataset

2.1. Dataset Composition

Fig. 8 list all question types divided along two dimensions. The horizontal dimension indicates the objects that need to be analyzed to answer the question successfully, and the vertical dimension indicates which facet of embodied scene understanding is evaluated. Detailed descriptions–along with two examples using both simulated and real observations–for each question type can be found at the end of this document in Sec. 3.3.2.

2.2. Zero-shot Answerability with Set-of-Mark Prompting

2.2.1. Human Evaluation

Before large-scale dataset generation, we first prepare a small questionnaire to examine the answerability and the quality of the MetaVQA Dataset. Since this is a pilot study, we utilize a Set-of-Mark prompting scheme slightly different from the final MetaVQA Dataset: contours are drawn around objects, and the background color of each label is determined by the corresponding text color following the original paper [5]. We sampled 35 questions with distinct types generated from a single keyframe to speed up the evaluation process. Six participants report an average accuracy of 88.05% on the 35 questions with a standard deviation of 7.54%. The best-performing participant achieves a 94.2& accuracy, while the worst-performing participant reports a 74.2% accuracy. An example question from the questionnaire is illustrated in Fig. 9.

Noticeably, participants struggle with question 19 (5 out of 6 wrong) and question 29 (4 out of 6 wrong), zero-indexed. The former is of type "order_leftmost", while the latter is of type "describe_distance." For question 19 illustrated in Fig. 10, the participants report–after questionnaire submission–confusion on whether the answer should be deduced using pixel-position ordering of the labels or the world-position ordering of objects. We speculate this

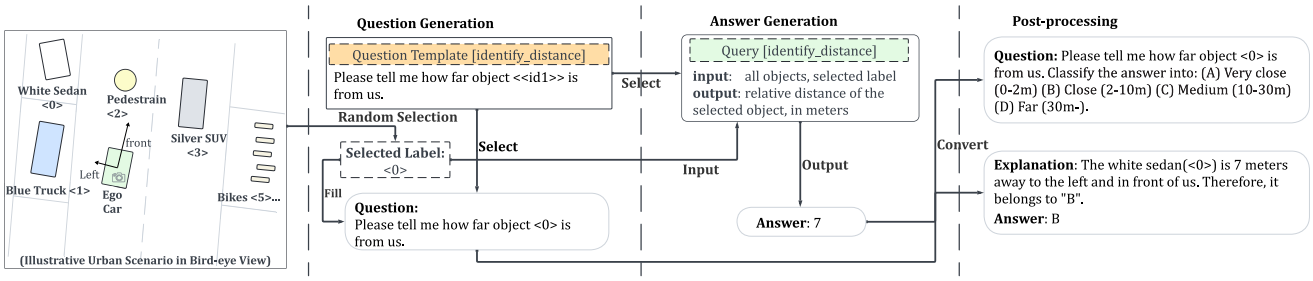


Figure 5. **Question-Answer generation of parameterized questions with only one type of parameter**.

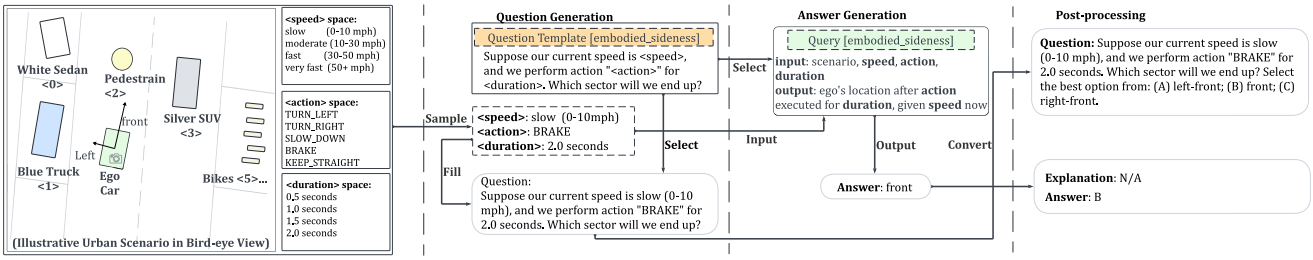


Figure 6. **Question-Answer generation of parameterized questions with distinct parameters**.

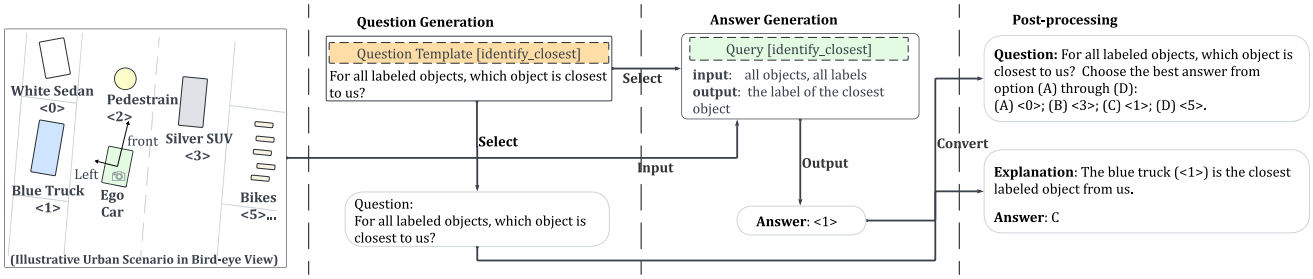


Figure 7. **Question-Answer generation of non-parameterized questions**.

confusion leads to the participants' overwhelming mistakes on this question. In addition, since this question involves objects very distant from the ego vehicle, the question is challenging due to the linear perspective. This might also cause conflicted participants' responses to question 29 shown in Fig. 11. Accounting for these factors, we refine the generation process for the final version of MetaVQA Dataset by choosing clearer phrasing and enforcing better visibility constraints on objects (for example, increasing the minimum observable pixels). Despite these issues, novice participants still report high test accuracies, and we conclude that the MetaVQA Dataset is intuitive to answer and clear in answering guidelines. Therefore, we argue that the MetaVQA Dataset is suitable for zero-shot plug-in-and-play evaluation of the embodied scene understanding entertained by general-purpose vision language models.

2.3. Effect of Set-of-Marks Prompting Scheme

The Set-of-Mark [5] paper proposes numerous prompting schemes, from using instance-segmentation masks to bounding boxes. In addition, the text size and background colors are also varied. We perform a grid search with observation generated using different prompting schemes while keeping the base images and object-to-label mapping identical across sets, and we use Qwen2 [6]–the VLM with the best grounding capability as discussed in the main paper. Referring to Tab. 1, Qwen2 achieves the best overall and grounding performance on images annotated with bounding boxes with labels of text size 1.25 and black for background colors. In addition, we observe that text size seems to have a trivial impact on the final performance. Based on these observations, we fixed the annotation style of MetaVQA with bounding-box annotations, black text background color, and a text size of 1.00 (to reduce label occlusions).

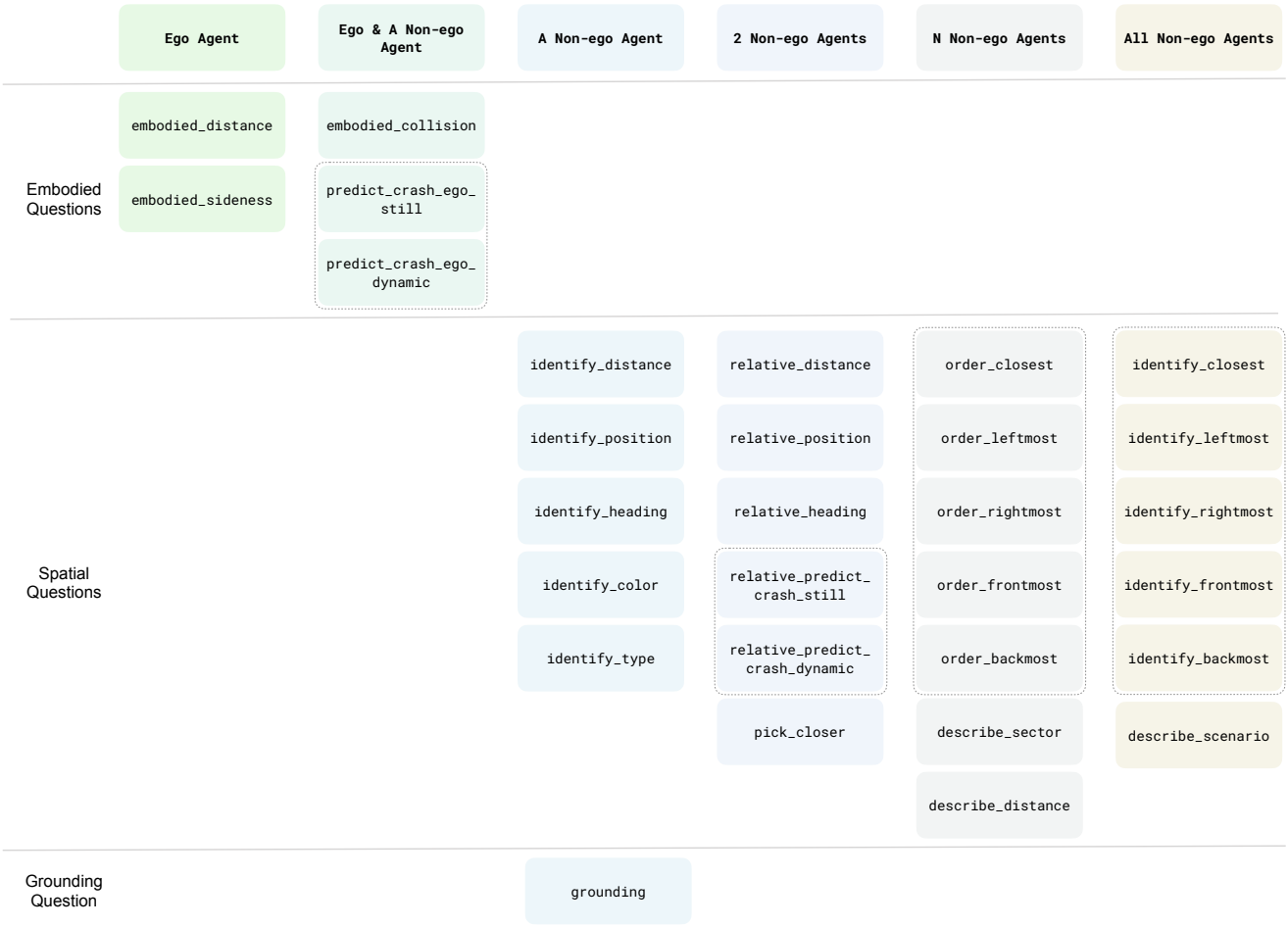


Figure 8. **Question taxonomy** of MetaVQA Dataset. Notice that questions are further blocked by black dotted contours to denote similarity in the formulation (illustrated collectively in Sec. 3.3.2).

Figure 9. **Sample question from the questionnaire**. The answer is (C).

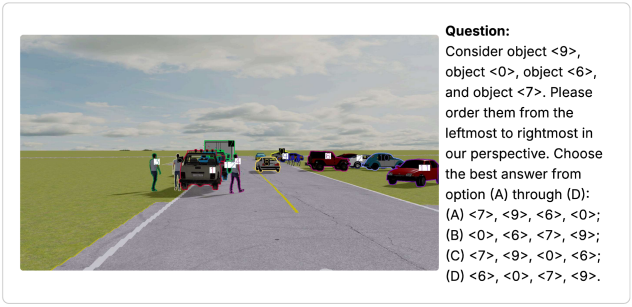


Figure 10. **Question 19 from the questionnaire**. The answer is (A). Ambiguous wording and distant objects lead to common mistakes by participants.

3. Benchmark Results

3.1. Definitions

We present the naming conventions used in this work in this subsection.

3.2. Visual Question Answering

We benchmark the performance of various baselines [6–10, 12] on the withheld test set ("overall") mentioned in the main paper. Furthermore, We provide detailed performances of baselines on (1) test questions with simulated

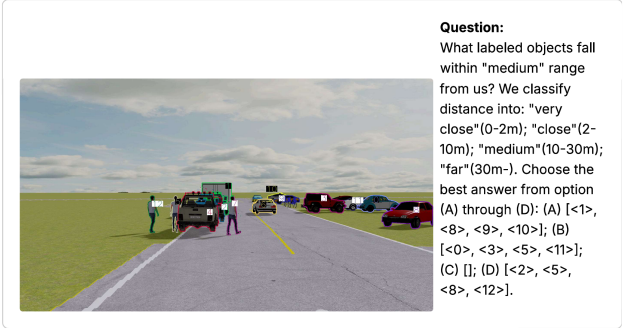


Figure 11. **Question 29 from the questionnaire**. The answer is (D). Some referred objects show limited visibility, leading to common errors.

Text Size	Form	Background	Overall	Grounding
0.75	box	white	0.440	0.867
0.75	box	black	0.457	**0.933**
0.75	mask	white	0.422	0.467
0.75	mask	black	0.420	0.533
0.75	contour	white	0.430	0.467
0.75	contour	black	0.420	0.733
1.25	box	white	0.437	0.800
1.25	box	black	**0.472**	**0.933**
1.25	mask	white	0.440	0.333
1.25	mask	black	0.437	0.333
1.25	contour	white	0.437	0.400
1.25	contour	black	0.422	0.600

Table 1. **Effect of Set-of-Marks Annotations.** We tested different annotation styles, text sizes, and background colors while fixing the model (Qwen2) and the numerical labeling, base images, and grounding questions.

Abbreviation	Checkpoint
LLaVA-NeXT	llava-1.6-vicuna-7b [7]
LLaVA-OneVision	llava-onevision-7b-ov [8]
GPT-4o	GPT-4o [9]
Qwen2	qwen2-vl-7b-instruct [6]
Llama3.2	llama-3.2-11B-Vision-Instruct [10]
InternVL2-4B	InternVL2-4B [11]
InternVL2-8B	InternVL2-8B [11]

Table 2. **Model Abbreviations**. These mappings are used consistently throughout the main paper and the supplementary materials.

observations ("sim" split) (2) test questions with real observations ("real" split). To save space, we used abbreviations illustrated in Tab. 2 for baselines in these benchmark tables.

3.2.1. Response Parsing

We establish a unified parsing standard using regular expression (regex) matching for the token sequences generated by all VLMs. If only a singular token is generated, we use this character as the option. If this is not the case, we search for option keywords provided in the multiple-choice questions. In cases of multiple matches, We select the last matched string as the model's output upon empirical examinations of the VLMs' raw outputs. If there is still no match, the parser will look for single characters enclosed by parentheses. If all searches return ill-composed results (empty match or illegal character), we consider the parsing to be a failed case. In the closed-loop evaluations, if a parse failure happens, a randomized action is taken. Code implementation will be available in the Github repository.

3.2.2. Benchmarks on Test Set

Tab. 3 presents the performance in "spatial reasoning" of the baseline VLMs on the withheld test set. Tab. 6 presents the performance in "embodied understanding" on the withheld test set. Tab. 7 presents the grounding performance of baseline VLMs, categorized according to the test set compositions.

3.2.3. Benchmarks on Real Test Split

Tab. 4 presents the performance in "spatial reasoning" of the baseline VLMs on the "real" split of withheld test set. Tab. 8 presents the performance in "embodied understanding" on the "real" split.

3.2.4. Benchmarks on Simulated Test Split

Tab. 5 presents the performance in "spatial reasoning" of the baseline VLMs on the "sim" split of withheld test set. Tab. 9 presents the performance in "embodied understanding" on the "sim" split.

3.3. Closed-loop Evaluation

3.3.1. Task Formulation

Interaction Paradigm. We use the MetaDrive [2] simulator, which provides accurate vehicle dynamics simulation for closed-loop evaluations. VLMs are deployed as driving agents in imported scenarios using [3]. At every five simulation steps (0.5 seconds wall time), the tested VLM is provided with (1) a Set-of-Mark annotated observation captured from the ego's front camera in 1600×900 resolution and (2) a driving prompt containing current navigation command and allowed discrete action space. The model will analyze the combined input and select the best action from available options. The chosen action will be fed into the simulation, and it will be repeated for the next 0.5 seconds (5 steps in simulation time) until the next inference step. Fig. 12 illustrates this process. The simulations terminate when their time horizons are reached or when the ego vehicle wanders off drivable regions.

Very rarely, the tested VLM will generate an invalid response according to the parser mentioned in Sec. 3.2.1. In this situation, we fix the chosen action as "KEEP_STRAIGHT" such that the speed and the heading of the ego vehicle will remain roughly identical.

Navigation command. At each inference step, the navigation command is recomputed to adjust for the current position of the ego vehicle. The command follows the following form:

Model	Spatial Questions (Overall)																								
	Overall	relative_distance	order_rightmost	describe_distance	identify_closest	relative_predict_crash_still	order_closest	identify_heading	identify_rightmost	relative_heading	relative_predict_crash_dynamic	identify_distance	order_leftmost	identify_type	order_backmost	identify_backmost	order_frontmost	relative_position	describe_sector	identify_frontmost	pick_closer	identify_position	identify_leftmost	identify_color
random	0.287	0.267	0.218	0.245	0.254	0.510	0.240	0.308	0.241	0.535	0.467	0.289	0.215	0.250	0.250	0.234	0.213	0.272	0.265	0.253	0.281	0.221	0.246	0.315
LLaVA-NeXT [7]	0.190	0.183	0.239	0.149	0.060	0.206	0.201	0.147	0.139	0.450	0.467	0.054	0.267	0.000	0.297	0.092	0.312	0.203	0.272	0.049	0.127	0.295	0.113	0.000
LLaVA-OneVision [8]	0.422	0.233	0.401	0.226	0.590	0.779	0.338	0.171	0.646	0.599	0.948	0.126	0.600	0.674	0.324	0.454	0.255	0.319	0.398	0.444	0.382	0.326	0.669	0.460
GPT-4o [9]	0.489	0.254	0.585	0.234	0.575	0.740	0.403	0.360	0.639	0.609	0.887	0.329	0.541	0.663	0.385	0.560	0.355	0.440	0.626	0.340	0.342	0.593	0.599	0.516
Qwen2 [6]	0.411	0.221	0.408	0.381	0.687	0.186	0.390	0.199	0.658	0.530	0.396	0.220	0.644	0.808	0.385	0.539	0.340	0.319	0.386	0.173	0.351	0.474	0.669	0.492
Qwen2-finetuned	0.740	0.550	**0.761**	0.891	0.843	0.627	**0.766**	0.404	0.899	0.540	0.821	0.487	**0.904**	**0.902**	**0.770**	**0.922**	**0.801**	0.431	**0.985**	0.759	0.325	**0.979**	**0.873**	0.831
Llama3.2 [10]	0.442	0.308	0.577	0.207	0.507	0.446	0.351	0.226	0.633	0.490	0.915	0.484	0.511	0.699	0.365	0.461	0.355	0.310	0.519	0.302	0.382	0.330	0.570	0.677
Llama3.2-finetuned	0.610	**0.667**	0.465	0.544	0.821	**0.873**	0.630	0.435	**0.905**	**0.688**	**0.953**	0.787	0.126	0.804	0.514	0.858	0.099	0.207	0.658	**0.772**	0.215	0.512	0.817	0.758
InternVL2-8B [12]	0.476	0.421	0.317	0.241	0.664	0.858	0.370	0.363	0.601	0.569	**0.953**	0.415	0.504	0.652	0.372	0.482	0.227	0.349	0.568	0.364	0.338	0.418	0.648	0.492
InternVL2-8B-finetuned	**0.813**	0.600	0.669	**0.904**	**0.866**	0.868	0.734	**0.647**	0.804	0.678	**0.953**	**0.834**	0.741	0.899	0.696	0.865	0.745	**0.776**	0.927	0.759	**0.794**	0.940	0.824	**0.863**

Table 3. **VQA benchmarks (Overall-Spatial)**. Per-question accuracies are evaluated on the withheld test set.

Model	Spatial Questions (Real)																							
	Overall	relative_distance	order_rightmost	describe_distance	identify_closest	relative_predict_crash_still	order_closest	identify_heading	identify_rightmost	relative_heading	relative_predict_crash_dynamic	identify_distance	order_leftmost	identify_type	order_backmost	identify_backmost	order_frontmost	relative_position	describe_sector	identify_frontmost	pick_closer	identify_position	identify_leftmost
random	0.296	0.237	0.188	0.250	0.203	0.500	0.362	0.317	0.194	0.602	0.490	0.279	0.210	0.280	0.246	0.236	0.185	0.287	0.256	0.310	0.286	0.245	0.244
LLaVA-NeXT	0.187	0.211	0.219	0.167	0.017	0.209	0.241	0.106	0.129	0.470	0.529	0.024	0.290	0.000	0.279	0.036	0.370	0.218	0.250	0.069	0.114	0.224	0.178
LLaVA-OneVision	0.452	0.228	0.406	0.294	0.542	0.764	0.466	0.174	0.694	0.614	**0.941**	0.091	0.629	0.826	0.246	0.364	0.185	0.366	0.494	0.362	0.381	0.476	0.778
GPT-4o	0.509	0.246	0.703	0.256	0.508	0.836	0.328	0.323	0.677	0.566	0.873	0.321	0.645	0.770	0.377	0.545	0.315	0.396	0.685	0.276	0.314	0.633	0.733
Qwen2	0.405	0.158	0.469	0.439	0.610	0.218	0.431	0.518	0.710	0.518	0.373	0.176	0.710	0.907	0.361	0.527	0.278	0.366	0.315	0.155	0.286	0.531	0.667
Qwen2-finetuned	0.723	0.649	**0.781**	**0.894**	0.780	0.591	0.759	0.342	**0.871**	0.651	0.745	0.515	**0.919**	**0.963**	0.770	0.891	**0.833**	0.406	**0.976**	0.621	0.295	**0.986**	0.844
Llama3.2	0.464	0.368	0.594	0.189	0.475	0.436	0.362	0.211	0.629	0.578	0.931	0.539	0.548	0.832	0.410	0.400	0.296	0.317	0.583	0.207	0.352	0.374	0.622
Llama3.2-finetuned	0.627	**0.728**	0.578	0.522	0.729	0.864	0.500	0.422	**0.871**	0.590	**0.941**	**0.824**	0.048	0.938	0.475	0.818	0.019	0.208	0.667	**0.810**	0.219	0.735	0.867
InternVL2-8B	0.516	0.283	0.328	0.283	0.644	**0.873**	0.379	0.675	0.694	0.675	**0.941**	0.424	0.548	0.795	0.410	0.400	0.278	0.376	0.708	0.259	0.305	0.503	0.711
InternVL2-8B-finetuned	**0.838**	0.640	0.672	0.878	**0.847**	0.864	**0.793**	**0.696**	0.790	**0.735**	**0.941**	0.640	0.823	0.950	**0.803**	**0.909**	0.759	**0.822**	0.958	**0.810**	**0.781**	0.952	**0.911**

Table 4. **VQA benchmarks (Real-Spatial)**. Per-question accuracies are evaluated on the "real" split of the withheld test set.

Model	Spatial Questions (Sim)																							
	Overall	relative_distance	order_rightmost	describe_distance	identify_closest	relative_predict_crash_still	order_closest	identify_heading	identify_rightmost	relative_heading	relative_predict_crash_dynamic	identify_distance	order_leftmost	identify_type	order_backmost	identify_backmost	order_frontmost	relative_position	describe_sector	identify_frontmost	pick_closer	identify_position	identify_leftmost	identify_color
random	0.281	0.294	0.244	0.242	0.293	0.521	0.167	0.298	0.271	0.487	0.445	0.304	0.219	0.209	0.253	0.233	0.230	0.260	0.270	0.221	0.276	0.196	0.247	0.315
LLaVA-NeXT	0.192	0.159	0.256	0.138	0.093	0.202	0.177	0.198	0.146	0.437	0.409	0.098	0.247	0.000	0.310	0.128	0.276	0.191	0.287	0.038	0.138	0.370	0.082	0.000
LLaVA-OneVision	0.398	0.238	0.397	0.185	0.627	0.798	0.260	0.168	0.615	0.588	0.955	0.179	0.575	0.461	0.379	0.512	0.299	0.282	0.332	0.490	0.382	0.167	0.619	0.460
GPT-4o	0.474	0.262	0.487	0.221	0.627	0.628	0.448	0.405	0.615	0.639	0.900	0.339	0.452	0.513	0.391	0.570	0.379	0.473	0.586	0.375	0.366	0.551	0.536	0.516
Qwen2	0.415	0.278	0.359	0.346	0.747	0.149	0.365	0.282	0.625	0.538	0.418	0.286	0.589	0.670	0.402	0.547	0.379	0.282	0.434	0.183	0.407	0.413	0.670	0.492
Qwen2-finetuned	0.754	0.460	**0.744**	0.889	**0.893**	0.670	**0.771**	0.481	0.917	0.462	0.891	0.446	**0.890**	0.817	**0.770**	**0.942**	**0.782**	0.450	**0.992**	**0.837**	0.350	**0.971**	**0.887**	0.831
Llama3.2	0.424	0.254	0.564	0.218	0.533	0.457	0.344	0.244	0.635	0.429	0.900	0.402	0.479	0.513	0.333	0.500	0.391	0.305	0.475	0.356	0.407	0.283	0.546	0.677
Llama3.2-finetuned	0.596	**0.611**	0.372	0.557	**0.893**	**0.883**	0.708	0.450	**0.927**	**0.756**	**0.964**	0.732	0.192	0.617	0.540	0.884	0.149	0.206	0.652	0.750	0.211	0.275	0.794	0.758
InternVL2-8B	0.444	0.413	0.308	0.215	0.680	0.840	0.365	0.450	0.542	0.496	**0.964**	0.402	0.466	0.452	0.345	0.535	0.195	0.328	0.471	0.423	0.366	0.326	0.619	0.492
InternVL2-8B-finetuned	**0.793**	0.563	0.667	**0.919**	0.880	0.872	0.698	**0.588**	0.813	0.639	**0.964**	**0.795**	0.671	**0.826**	0.621	0.837	0.736	**0.740**	0.906	0.731	**0.805**	0.928	0.784	**0.863**

Table 5. **VQA benchmarks (Sim-Spatial)**. Per-question accuracies are evaluated on the "sim" split of the withheld test set.

`your final destination is at <distance> to <position> at this moment.`

Here, the `<distance>` and `<position>` parameters will be replaced with concrete values chosen from the discrete vocabulary for spatial information mentioned in Sec. 1.3.

Action space. The actions in the driving prompts are statically mapped to low-level control signals to MetaDrive. MetaDrive receives normalized action as input to control the ego vehicle: $\mathbf{a} = [a_1, a_2]^T \in [-1, 1]^2$. At each simulation time step, MetaDrive converts the normalized action into the steering u_s (degree), acceleration u_a (hp) and brake signal u_b (hp) in the following ways: (i) $u_s = S_{max} a_1$, (ii) $u_a = F_{max} \max(0, a_2)$, (iii) $u_b = -B_{max} \min(0, a_2)$,

Model	Embodied Questions (Overall)					
	Overall	embodied_distance	embodied_collision	predict_crash ego_still	embodied_sideness	predict_crash ego_dynamic
random	0.382	0.255	0.498	0.521	0.348	0.504
LLaVA-NeXT	0.419	0.159	0.489	0.303	0.652	0.384
LLaVA-OneVision	0.746	0.442	0.923	0.976	0.794	**0.961**
GPT-4o	0.764	0.785	0.719	0.893	0.732	0.873
Qwen2	0.649	0.451	0.836	0.259	0.804	0.482
Qwen2-finetuned	**0.948**	**0.998**	0.879	0.817	**1.000**	0.894
Llama3.2	0.536	0.332	0.650	0.517	0.574	0.849
Llama3.2-finetuned	0.944	0.962	0.846	**0.997**	0.999	**0.961**
InternVL2-8B	0.711	0.620	0.923	0.914	0.509	**0.961**
InternVL2-8B-finetuned	0.926	0.807	**0.953**	**0.997**	**1.000**	**0.961**

Table 6. **VQA benchmarks (Overall-Embodied)**. Per-question accuracies are evaluated on the withheld test set.

Model	Grounding Questions		
	Overall	Real	Sim
random	0.268	0.257	0.280
LLaVA-NeXT	0.248	0.229	0.271
LLaVA-OneVision	0.728	0.827	0.615
GPT4-o	0.831	0.888	0.766
Qwen2	0.874	0.859	0.890
Qwen2-finetuned	**0.972**	**0.992**	**0.950**
Llama3.2	0.790	0.855	0.716
Llama3.2-finetuned	0.923	0.944	0.899
InternVL2-8B	0.702	0.783	0.610
InternVL2-8B-finetuned	0.916	0.948	0.881

Table 7. **VQA benchmarks (Grounding)**. Per-question accuracies are evaluated on the withheld whole test set, "real" split of the test set, and "sim" split of the test set.

Model	Embodied Questions (Real)					
	Overall	Emb_Dist	Emb_Coll	PredCrash EgoStill	Emb_Side	PredCrash EgoDyn
random	0.372	0.248	0.498	0.487	0.345	0.467
LLaVA-NeXT	0.414	0.189	0.445	0.342	0.647	0.327
LLaVA-OneVision	0.735	0.430	0.905	0.980	0.795	**0.980**
GPT-4o	0.762	0.784	0.706	0.895	0.739	0.873
Qwen2	0.653	0.446	0.852	0.322	0.796	0.453
Qwen2-finetuned	0.946	**0.999**	0.875	0.789	**1.000**	0.887
Llama3.2	0.542	0.339	0.654	0.566	0.580	0.867
Llama3.2-finetuned	**0.947**	0.969	0.844	**0.993**	0.999	**0.980**
InternVL2-8B	0.720	0.615	0.905	0.895	0.576	**0.980**
InternVL2-8B-finetuned	0.919	0.780	**0.956**	**0.993**	**1.000**	**0.980**

Table 8. **VQA benchmarks (Real-Embodied)**. Per-question accuracies are evaluated on the "real" split of the withheld test set. Question types are shortened for formatting.

Model	Embodied Questions (Sim)					
	Overall	Emb_Dist	Emb_Coll	PredCrash EgoStill	Emb_Side	PredCrash EgoDyn
random	0.395	0.264	0.497	0.558	0.353	0.545
LLaVA-NeXT	0.426	0.12	0.542	0.261	0.660	0.448
LLaVA-OneVision	0.760	0.457	0.945	0.971	0.793	**0.940**
GPT-4o	0.767	0.786	0.734	0.891	0.722	0.873
Qwen2	0.645	0.457	0.817	0.188	0.815	0.515
Qwen2-finetuned	**0.949**	**0.998**	0.883	0.848	**1.000**	0.903
Llama3.2	0.527	0.321	0.644	0.464	0.565	0.828
Llama3.2-finetuned	0.939	0.952	0.847	**1.000**	**1.000**	**0.940**
InternVL-8B	0.699	0.627	0.945	0.935	0.418	**0.940**
InternVL-8B-finetuned	0.936	0.843	**0.949**	**1.000**	**1.000**	**0.940**

Table 9. **VQA benchmarks (Sim-Embodied)**. Per-question accuracies are evaluated on the "sim" split of the withheld test set. Question types are shortened for formatting.

wherein S_{max} (degree) is the maximal steering angle, F_{max} (hp) is the maximal engine force, and B_{max} (hp) is the maximal brake force. For fair and replicable experiments, we use identical vehicle configurations(for example, maximum engine force) across different trials.

We conducted grid searches to fix the suitable set of actions. For each candidate, we reconstruct real-world driving trajectories as action sequences with only allowed action provided by the candidate. These sequences are computed greedily (and repeated) at every five simulation steps, following the same inference frequency as the closed-loop evaluation. The optimal action at a particular step is decided according to the resulting deviation from the original trajectories if the action is executed. This sequence-building is autoregressive, meaning that previous optimal actions(and their generated trajectories) affect the decision on later optimal actions. We fix the current action space as it leads to the best reconstruction quality.

Test scenarios. We tailor 120 diverse scenarios to evaluate VLMs' embodied scene understanding holistically. These scenarios include 60 from the nuScenes dataset and the other 60 selected from a corpus of safety-critical situations generated using CAT [13]. For each of the 60 safety-critical scenarios, an adversarial agent will attempt to run into the ego vehicle, and we ensure the observability of adversarial agents.

3.3.2. Metrics

Route Completion The ratio of the traveled distance against the length of the complete route averaged across scenarios.

Collision Rate The ratio of scenarios where the ego vehicle collides with any other object.

Off-Road Rate The ratio of scenarios where the ego vehicle leaves drivable regions.

Final Displacement Error (FDE) The L2 distance between the final position of the ego vehicle from the final destination averaged across scenarios.

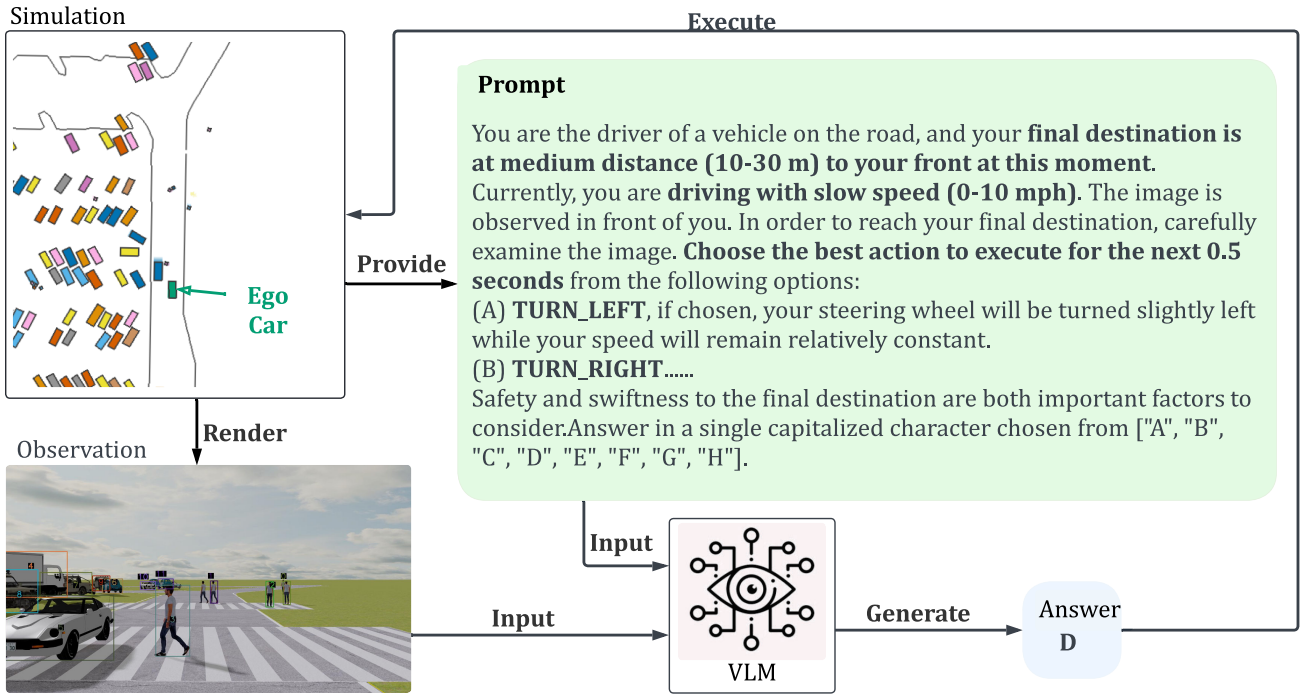


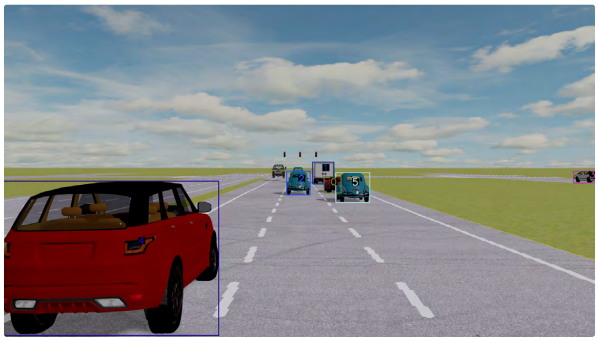
Figure 12. **Closed-loop evaluation paradigm.**

Average Displacement Error (ADE) The mean per-step L2 distance between the ground-truth trajectories and the VLM-driven trajectories averaged across scenarios. If a simulation terminates prematurely(due to VLMs driving off-road), the last ego vehicle position is appended to align the length of the ground-truth trajectory with the VLM-driven trajectory.

Embodied Questions:

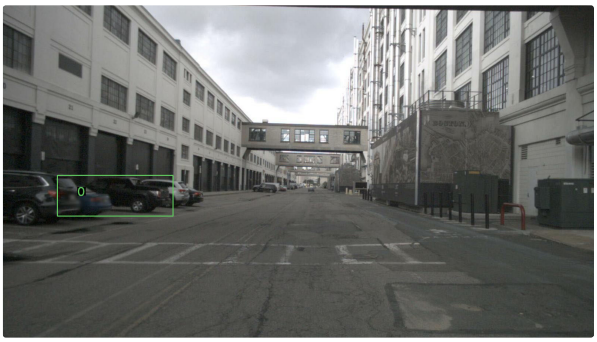
embodied_distance. This question examines how far the ego will move from the current position, assuming that <action> is executed over the next <duration> period and the ego's current speed is <speed>.

Question:
Suppose our current speed is fast(30-50 mph), and we perform action "BRAKE" for 1.5 seconds. How far will we end up from our current position? Select the best option from: (A) Very close(0-2m); (B) Close(2-10m); (C) Medium(10-30m); (D) Far(30m-)
Explanation: N/A
Answer: C

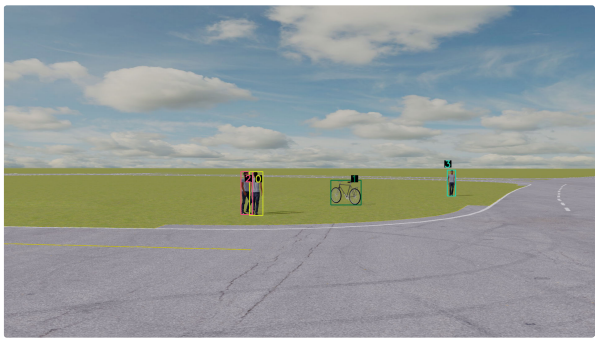


Question:
Suppose our current speed is moderate(10-30 mph), and we perform action "SLOW_DOWN" for 1.0 seconds. How far will we end up from our current position? Select the best option from:(A) Very close(0-2m); (B) Close(2-10m); (C) Medium(10-30m); (D) Far(30m-)
Explanation: N/A
Answer: B

embodied_sideness. This question examines how whether the ego will move to its left or its right(in the current frame), assuming that <action> is executed over the next<duration> period and the ego's current speed is <speed>.



Question:
Suppose our current speed is moderate (10-30 mph), and we perform action "KEEP_STRAIGHT" for 1.0 seconds. Which sector will we end up? Select the best option from: (A) left-front; (B) front; (C) right-front.
Explanation: N/A
Answer: B



Question:
Suppose our current speed is moderate (10-30 mph), and we perform action "TURN_LEFT" for 1.5 seconds. Which sector will we end up? Select the best option from: (A) left-front; (B) front; (C) right-front.

Explanation: N/A
Answer: A

Embodied Questions:

embodied_collision. This question examines whether the ego will collide into selected object <id1>, assuming that <action> is executed over the next <duration> period and the ego's current speed is <speed>.

Question:
Suppose our current speed is slow(0-10 mph), and we perform action "BRAKE" for 0.5 seconds. Will we run into object <0>, provided that it remains still? Select the best option from: (A) Yes; (B) No.
Explanation:
We will not run into object <0>, even though we both end in our front sector.
Answer: B

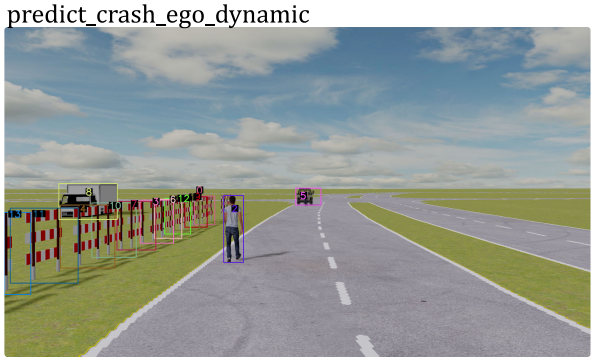
Question:
Suppose our current speed is fast(30-50 mph), and we perform action "SLOW_DOWN" for 1.0 seconds. Will we run into object <0>, provided that it remains still? Select the best option from: (A) Yes; (B) No.
Explanation:
We will not run into object <0>. Object <0> is located in the right-front sector, but we will end in the front sector.
Answer: B

predict_crash_ego_*. This family of questions examines how whether the selected object <id1> will collide with the ego under various conditions.

predict_crash_ego_still

Question:
Suppose object <0> proceed along its current heading. Will it collides into us if we stay still? Choose the best answer between option (A) and (B): (A) Yes; (B) No.
Explanation:
No, this bus (<0>) to the right and in front of us and heading toward our front direction will not run into us if it drives along its current heading.
Answer: B

predict_crash_ego_dynamic



Question:
Suppose both object <5> and us proceed along our corresponding current headings with the same speed. Will we collide into each other? Choose the best answer between option (A) and (B): (A) Yes; (B) No.
Explanation:
No, this gray pickup (<5>) directly in front of us and heading toward our front direction will not run into us.
Answer: B

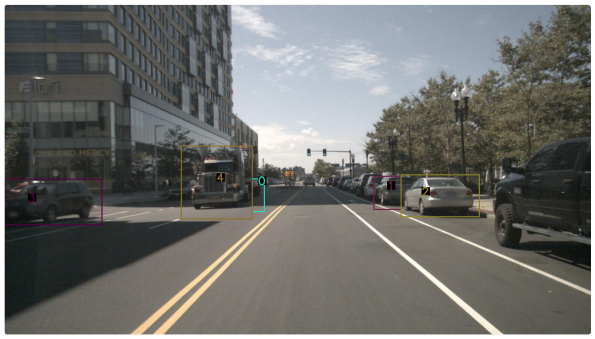
Spatial Questions:

identify_distance. This question prompts VLMs to estimate the distance of the selected object <id1> from the ego.

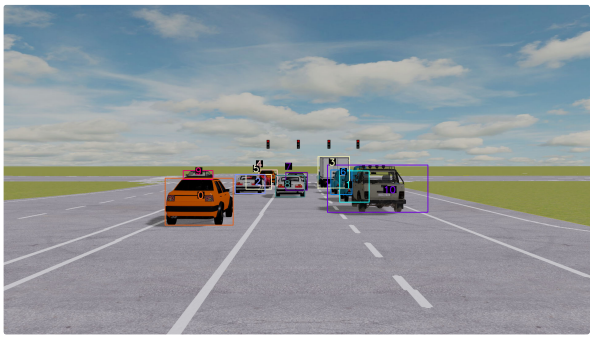
Question:
Please tell me how far object <0> is from us. Classify the answer into: (A) Very close (0-2m) (B) Close (2-10m) (C) Medium (10-30m) (D) Far(30m-).
Explanation:
The truck (<0>) is 28 meters to the left and in front of us. Therefore, it belongs to "medium".
Answer: C

Question:
Please tell me how far object <4> is from us. Classify the answer into: (A) Very close (0-2m) (B) Close (2-10m) (C) Medium (10-30m) (D) Far (30m-).
Explanation:
The gray pickup (<4>) is 51 meters to the left and in front of us. Therefore, it belongs to "far".
Answer: D

identify_position. This question prompts VLMs to estimate the direction of the selected object <id1> from the ego.



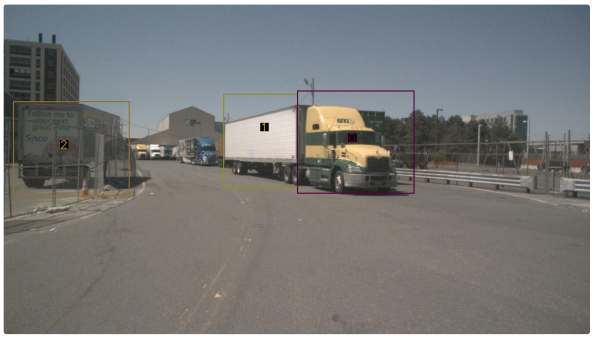
Question:
Please tell me the relative position of <3> with respect to us. Choose the best answer from option (A) through (D): (A) next-to; (B) back; (C) left-front; (D) right-front.
Explanation:
The car (<3>) is to the left and in front of us.
Answer: C



Question:
Please tell me the relative position of <8> with respect to us. Choose the best answer from option (A) through (D): (A) right-front; (B) front; (C) next-to; (D) left.
Explanation:
The white sedan (<8>) is directly in front of us.
Answer: B

Spatial Questions:

identify_heading. This question prompts models to estimate the heading angle of the selected object <id1>, expressed relative to the ego's front direction. The provided options are sufficiently distinct to avoid ambiguity.



Question:
Please describe the heading direction of object <0>. Choose the best answer from option (A) through (D): (A) left-front; (B) right-front; (C) right-back; (D) left-back.
Explanation:
The truck (<0>) directly in front of us is facing our right-back direction.
Answer: C

Question:
Please describe the heading direction of object <1>. Choose the best answer from option (A) through (D): (A) front; (B) right; (C) left; (D) back.
Explanation:
The red sports car (<1>) to the left and in front of us is facing our right direction.
Answer: B

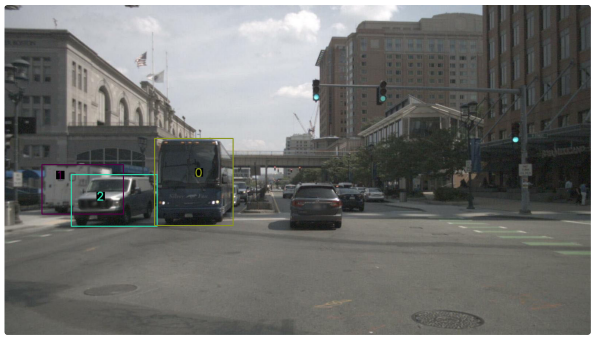
identify_color This question prompts models to select the color of object <id1>. Note that it is generated only with simulated observations, as "color" is not annotated in the nuScenes dataset.

Question:
Specify the color of object <1>. Choose the best answer from option (A) through (D): (A) White; (B) Black; (C) Grey; (D) Yellow.
Explanation:
The color of this hatchback (<1>) to the right and in front of us is yellow.
Answer: D

Question:
Specify the color of object <1>. Choose the best answer from option (A) through (D): (A) White; (B) Black; (C) Grey; (D) Yellow.
Explanation:
The color of this hatchback (<1>) to the right and in front of us is yellow.
Answer: D

Spatial Questions:

identify_type. This question prompts VLMs to select the most descriptive type of the selected object <id1>.

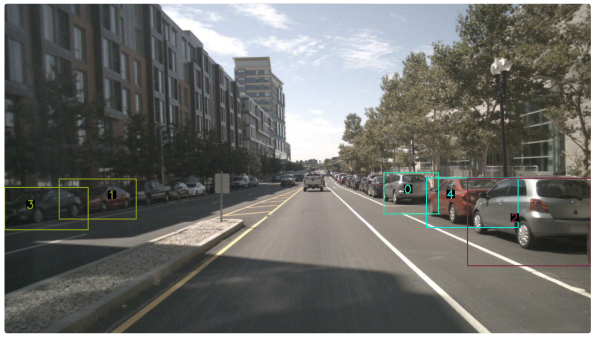


Question:
Specify the type of object <0>. Choose the best answer from option (A) through (D): (A) Car; (B) Hatchback; (C) Truck; (D) Bus.
Explanation:
The type of this object(<0>) to the left and in front of of us is "bus".

Answer: D

Question:
Specify the type of object <4>. Choose the best answer from option (A) through (D): (A) Hatchback; (B) Sedan; (C) Suv; (D) Sports Car.
Explanation:
The type of this white object(<4>) to the left and in front of of us is "sedan".
Answer: B

relative_distance. This question prompts VLMs to select the relative distance between two objects <id1> and <id2>.

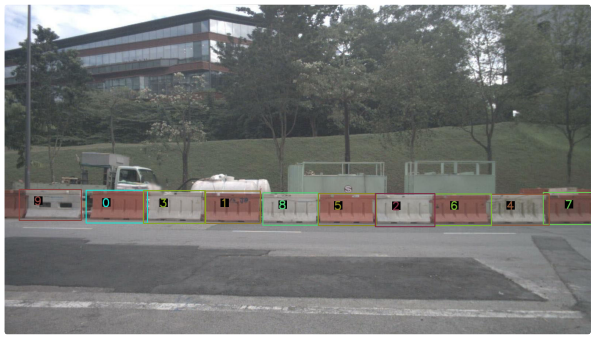


Question:
How close are object <1> and object <3> positioned? Classify the answer into: (A) Very close(0-2m); (B) Close(2-10m); (C) Medium(10-30m); (D) Far(30m-).
Explanation:
Object <1>, a car to the left and in front of us, is directly in front of object <3>, a car to the left and in front of us, at a close distance.
Answer: B

Question:
How close are object <7> and object <0> positioned? Classify the answer into: (A) Very close(0-2m); (B) Close(2-10m); (C) Medium(10-30m); (D) Far(30m-).
Explanation:
Object <7>, a gray pickup to the right and in front of us, is to the right of object <0>, a yellow hatchback to the left and in front of us, at a medium distance.
Answer: C

Spatial Questions:

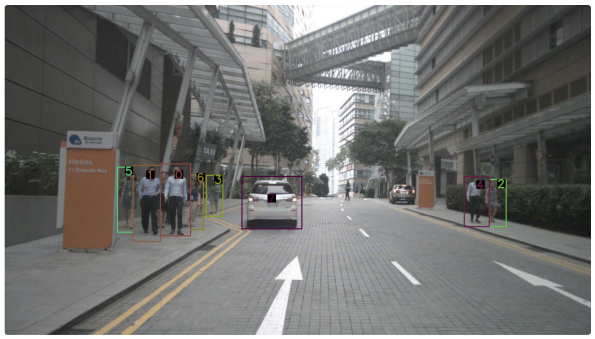
relative_position. This question prompts VLMs to evaluate how is object <id1> related spatially with object <id12>, expressed in the ego perspective.



Question:
How is object <6> positioned relative to object <2>? Choose the best answer from option (A) through (D): (A) next-to; (B) right-back; (C) right; (D) left-front.
Explanation:
Object <6>, a traffic barrier to the right and in front of us, is to the right of object <2>, a traffic barrier to the right and in front of us.
Answer: C

Question:
How is object <4> positioned relative to object <7>? Choose the best answer from option (A) through (D): (A) right-back; (B) right-front; (C) next-to; (D) left-back.
Explanation:
Object <4>, a blue hatchback to the right and in front of us, is to the right and in front of object <7>, a gray sedan directly in front of us.
Answer: B

relative_heading. This question prompts VLMs to determine if object <id1> and <id12> are heading towards roughly the same direction.



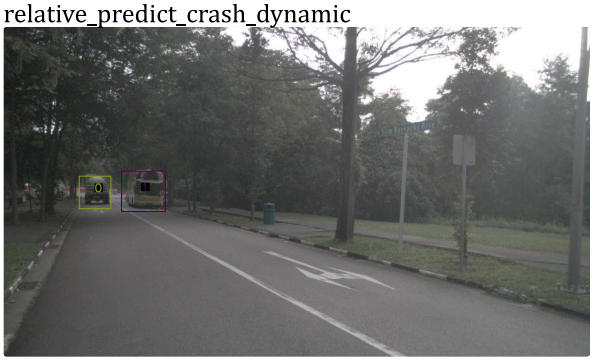
Question:
Is object <7> heading toward roughly the same direction as object <1>? Choose the best answer between option (A) and (B): (A) Yes; (B) No.
Explanation:
No. Object <7>, a car directly in front of us, is not heading toward the same direction as object <1>, a pedestrian located to the left and in front of us. In particular, object <1>'s heading differs by 194 degrees counterclockwise from that of object <7>.
Answer: B

Question:
Is object <2> heading toward roughly the same direction as object <1>? Choose the best answer between option (A) and (B): (A) Yes; (B) No.
Explanation:
No. Object <2>, a blue hatchback to the right and in front of us, is not heading toward the same direction as object <1>, a white bike located directly in front of us. In particular, object <1>'s heading differs by -81 degrees counterclockwise from that of object <2>.
Answer: B

Spatial Questions:

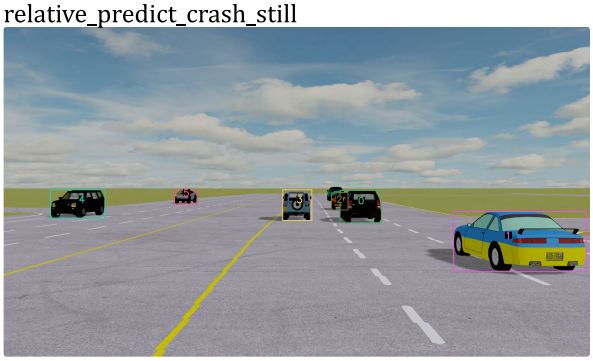
relative_predict_crash_*. This family of questions prompts VLMs to infer whether two objects <id1> and <id12> will collect under varying assumptions.

relative_predict_crash_dynamic



Question:
Suppose object <0> and object <1> proceed along their current directions with the same speed. Will they collide into each other? Choose the best answer between option (A) and (B): (A) No; (B) Yes.
Explanation:
No, object <0> will not run into object <1>.
Answer: A

relative_predict_crash_still



Question:
Suppose object <1> proceed along its current heading. Will it collides into object <0> if object <0> stays still? Choose the best answer between option (A) and (B): (A) No; (B) Yes.
Explanation:
Yes, object <1> will run into object <0>.
Answer: B

pick_closer. This question asks the VLM to select the closer object from two candidates.

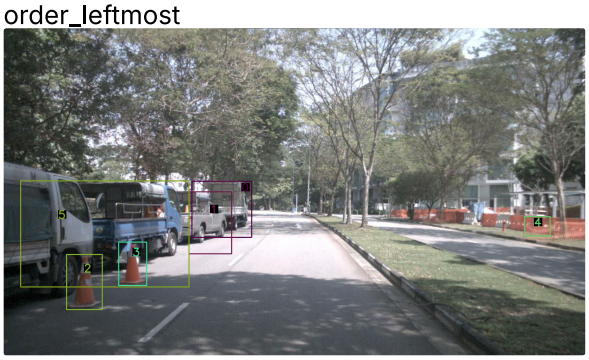
Question:
Which object is closer to me, <4> or <3>? Choose the best answer from option (A) through (C): (A) <4> is closer; (B) <3> is closer; (C) <4> and <3> are about the same distance.
Explanation:
Object <3>, a truck to the right and in front of us, is closer to us than object <4>, a traffic barrier directly in front of us.
Answer: C

Question:
Which object is closer to me, <2> or <11>? Choose the best answer from option (A) through (C): (A) <2> is closer; (B) <11> is closer; (C) <2> and <11> are about the same distance.
Explanation:
Object <11>, a yellow sports car to the right of us, is closer to us than object <2>, a white sedan to the left and in front of us.
Answer: C

Spatial Questions:

order_*st. This family of questions asks the VLM to attend to multiple objects and sort their relevance by some spatial ordering in top-down world coordinates.

order_leftmost



Question:
Consider object <5>, object <2>, object <3>, and object <0>. Please order them from leftmost to rightmost in our coordinate system. Choose the best answer from option (A) through (D):
(A) <3>, <0>, <2>, <5>
(B) <5>, <3>, <2>, <0>
(C) <0>, <5>, <3>, <2>
(D) <2>, <3>, <5>, <0>
Explanation:
The truck(<0>) is at the far left, and the traffic cone(<2>) is at the far right. The truck(<5>) and the traffic cone(<3>) are in between.
Answer: C

order_frontmost

Question:
Consider object <0>, object <2>, object <5>, and object <4>. Please order them from furthest to the closest along our front direction. Choose the best answer from option (A) through (D):
(A) <5>, <0>, <2>, <4>
(B) <2>, <4>, <0>, <5>
(C) <4>, <5>, <0>, <2>
(D) <2>, <5>, <0>, <4>
Explanation:
The white sedan(<5>) is at the furthest along our heading direction, and the black sedan(<4>) is the closest. The yellow hatchback(<0>) and the white sedan(<2>) are in between.
Answer: A

describe_sector. This question asks the VLM to attend to all observable objects and select the maximal object set such that all of its members are in the specified ego's direction from the question body.

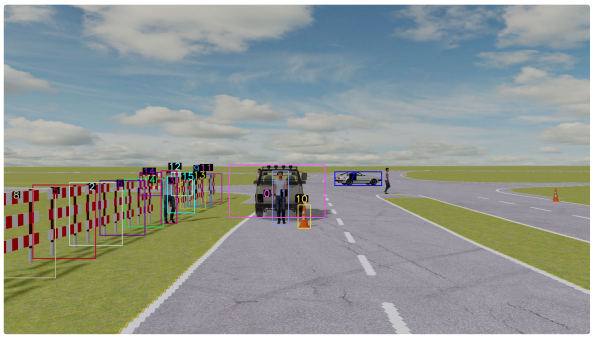
Question:
What labeled objects fall into our left-front sector? Choose the best answer from option (A) through (D):
(A) [ ]
(B) [<0>, <2>, <7>, <10>]
(C) [<1>, <3>, <7>, <10>]
(D) [<3>, <5>, <7>, <8>]
Explanation:
Option (A) is wrong since there exists at least 4 objects (<0>, <2>, <7>, <10>) in the specified(left-front) sector;
Option (C) is wrong since there exists 2 objects (<1>, <3>) in the right-front sector;
Option (D) is wrong since there exists 3 objects (<3>, <5>, <8>) in the right-front sector.
Answer: B

Question:
What labeled objects fall into our front sector? Choose the best answer from option (A) through (D):
(A) [ ]
(B) [<3>, <4>, <8>]
(C) [<3>, <4>, <9>]
(D) [<5>, <9>, <10>]
Explanation:
Option (A) is wrong since there exists at least 3 objects (<5>, <9>, <10>) in the specified(front) sector;
Option (B) is wrong since there exists 3 objects (<3>, <4>, <8>) in the left-front sector;
Option (C) is wrong since there exists 2 objects (<3>, <4>) in the left-front sector.
Answer: D

Spatial Questions:

describe_distance. This question asks VLMs to attend to all observable objects and select the maximal object set such that all of its members are located away from the ego by the specified distance from the question body.

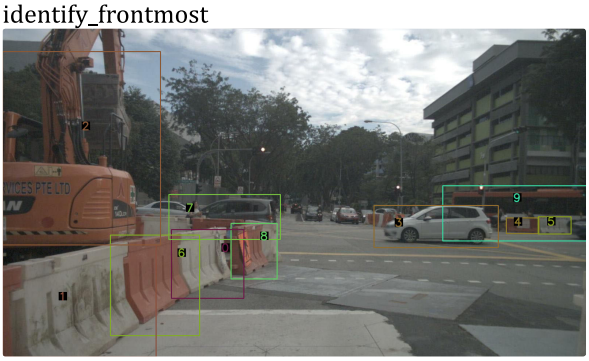
Question:
What labeled objects fall within "medium" range from us? We classify distance into: "very close"(0-2m); "close"(2-10m); "medium"(10-30m); "far"(30m-). Choose the best answer from option (A) through (D):
(A) [<0>, <1>, <2>]
(B) [<1>, <2>, <3>]
(C) [ ]
(D) [<0>, <2>, <3>]
Explanation:
Option (A) is wrong since there exists 1 object (<1>) positioned at far distance from us; Option (B) is wrong since there exists 1 object (<1>) positioned at far distance from us; Option (C) is wrong since there exists at least 3 objects (<0>, <2>, <3>) positioned at specified(medium) distance from us.
Answer: D



Question:
What labeled objects fall within "medium" range from us? We classify distance into: "very close"(0-2m); "close"(2-10m); "medium"(10-30m); "far"(30m-). Choose the best answer from option (A) through (D):
(A) [<1>, <5>, <10>, <15>]
(B) [ ]
(C) [<0>, <8>, <11>, <15>]
(D) [<5>, <9>, <13>, <14>]
Explanation:
Option (B) is wrong since there exists at least 4 objects (<1>, <5>, <10>, <15>) positioned at specified(medium) distance from us; Option (C) is wrong since there exists 1 object (<11>) positioned at far distance from us; Option (D) is wrong since there exists 1 object (<14>) positioned at far distance from us.
Answer: A

identify_*st. This question asks the VLM to attend to all observable objects and select the leading object according to some ordering specified in the question body.

identify_frontmost



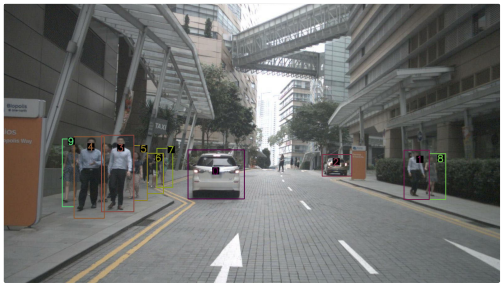
Question:
For all labeled objects, which object is furthest along our front direction? Choose the best answer from option (A) through (D): (A) <3>; (B) <4>; (C) <5>; (D) <9>.
Explanation:
The bus (<9>) is the object furthest along front direction us.
Answer: D

identify_closest

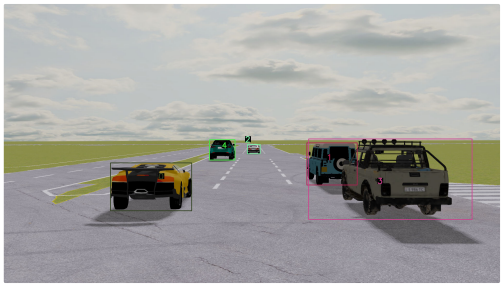
Question:
For all labeled objects, which object is closest to us? Choose the best answer from option (A) through (D): (A) <1>; (B) <0>; (C) <9>; (D) <3>.
Explanation:
The gray pickup (<1>) is the closest labeled object from us.
Answer: A

Spatial Questions:

describe_scenario. This question prompts the VLM to examine all labeled objects in the scenario. It is a train-only question designed to boost learning performance and avoid VLM collapse.



Question:
Please describe all labeled objects in the observation, in the order of their labels...
Explanation:
<0>: A car positioned in our front sector at medium distance. It heads toward our front direction.
<1>: A pedestrian positioned in our right-front sector at medium distance. It heads toward our back direction.
<2>: A car positioned in our right-front sector at far distance. It heads toward our front direction.
<3>: A pedestrian positioned in our left-front sector at medium distance. It heads toward our right-back direction.
<4>: A pedestrian positioned in our left-front sector at medium distance. It heads toward our right-back direction.
<5>: A pedestrian positioned in our left-front sector at medium distance. It heads toward our back direction.
<6>: A pedestrian positioned in our left-front sector at medium distance. It heads toward our right-back direction.
<7>: A pedestrian positioned in our left-front sector at medium distance. It heads toward our back direction.
<8>: A pedestrian positioned in our right-front sector at medium distance. It heads toward our back direction.
<9>: A pedestrian positioned in our left-front sector at medium distance. It heads toward our right-back direction.
Answer: N/A



Question:
Please describe all labeled objects in the observation, in the order of their labels...
Explanation:
<0>: A yellow sports car positioned in our left-front sector at medium distance. It heads toward our front direction.
<1>: A blue SUV positioned in our right-front sector at medium distance. It heads toward our front direction.
<2>: A white sedan positioned in our front sector at far distance. It heads toward our front direction.
<3>: A gray pickup positioned in our right-front sector at medium distance. It heads toward our front direction.
<4>: A blue SUV positioned in our left-front sector at far distance. It heads toward our front direction.
Answer: N/A

Grounding Questions:

grounding. This question examines the visual grounding ability of the tested VLM. All non-answer options are selected from valid labels to challenge the model maximally.

Question:
What is the numerical label associated with the highlighted area? Choose the best answer from option (A) through (D): (A) 1; (B) 26;(C) 30; (D) 28.
Explanation: N/A
Answer: B

Question:
What is the numerical label associated with the highlighted area? Choose the best answer from option (A) through (D): (A) 10; (B) 25; (C) 35; (D) 29.
Explanation: N/A
Answer: D

References

[1] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yuxin Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 1

[2] Quanyi Li, Zhenghao Peng, Lan Feng, Qihang Zhang, Zhenghai Xue, and Bolei Zhou. Metadrive: Composing diverse driving scenarios for generalizable reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1, 6

[3] Quanyi Li, Zhenghao Peng, Lan Feng, Zhizheng Liu, Chenda Duan, Wenjie Mo, and Bolei Zhou. Scenarionet: Open-source platform for large-scale traffic scenario simulation and modeling. *Advances in Neural Information Processing Systems*, 2023. 1, 6

[4] Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R Qi, Yin Zhou, et al. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9710–9719, 2021. 1

[5] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*, 2023. 2, 3, 4

[6] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2, 4, 5, 6, 7

[7] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 6, 7

[8] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 6, 7

[9] OpenAI. Chatgpt-4, 2024. Large Language Model developed by OpenAI. 6, 7

[10] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Mon-

talvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The llama 3 herd of models, 2024. 5, 6, 7

[11] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 6

[12] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023. 5, 7

[13] Linrui Zhang, Zhenghao Peng, Quanyi Li, and Bolei Zhou. Cat: Closed-loop adversarial training for safe end-to-end driving. In *7th Annual Conference on Robot Learning*, 2023. 8