Retrieval-Based Interleaved Visual Chain-of-Thought in Real-World Driving Scenarios

Charles Corbière* Simon Roburin* Syrielle Montariol*
Antoine Bosselut Alexandre Alahi

EPFL, Switzerland

Abstract

While chain-of-thought (CoT) prompting improves reasoning in large language models, its effectiveness in vision-language models (VLMs) remains limited due to over-reliance on textual cues and memorized knowledge. To investigate the visual reasoning capabilities of VLMs in complex real-world scenarios, we introduce DRIVINGVQA*, a visual question answering dataset derived from driving theory exams, which contains 3,931 multiple-choice problems with expert-written explanations and grounded entities relevant to the reasoning process. Leveraging this dataset, we propose RIV-CoT, a Retrieval-Based Interleaved Visual Chain-of-Thought method that enables VLMs to reason using visual crops corresponding to these relevant entities. Our experiments demonstrate that RIV-CoT improves answer accuracy by 3.1% and reasoning accuracy by 4.6% over vanilla CoT prompting. Furthermore, we demonstrate that our method effectively scales to the larger A-OKVQA reasoning dataset by leveraging automatically generated pseudo-labels, outperforming CoT prompting. Code and dataset are available at https://vita-epfl.github.io/DrivingVQA.*

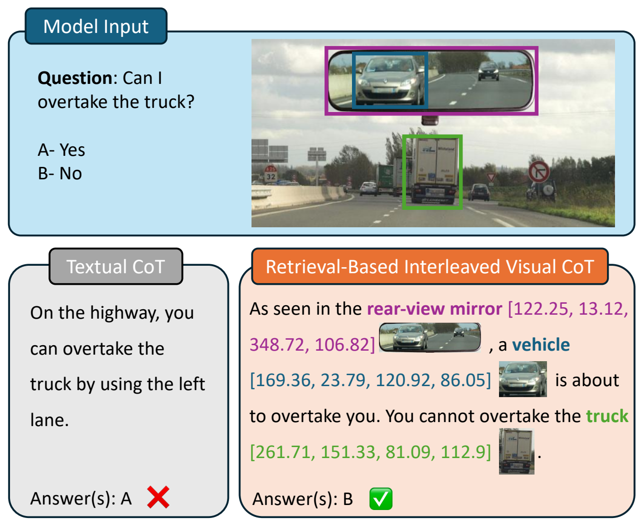


Figure 1. **Illustration of retrieval-based interleaved visual chain-of-thought in DRIVINGVQA**. Successfully answering the question requires detecting relevant entities (e.g., the truck, the car in the rear-view mirror), recognizing their attributes (e.g., the car signaling to overtake), and reasoning spatially to determine whether overtaking is safe. The interleaved explanation provides step-by-step reasoning aligned with visual content.

1. Introduction

Chain-of-thought (CoT) [51] is a prompting strategy that aims at enhancing the reasoning capabilities of large language models (LLMs) [13, 33, 46] and, more recently, vision-language models (VLMs) [1, 21, 23, 47]. While well-suited for mathematical and logical reasoning, CoT has limited effectiveness when it comes to visual, spatial, and multi-step reasoning [41, 44, 56].

A critical challenge is VLMs' tendency to excessively rely on textual inputs and memorized knowledge rather than visual inputs, which can lead to hallucinations [3, 9, 14, 24]. Recent methods attempt to enforce explicit grounding by generating textual image descriptions [20, 35, 57], leveraging scene graphs [32], or using bounding box coordinates [4, 16, 27]. However, these methods have limited expressiveness; textual descriptions or sets of bounding box coordinates are poorly leveraged by VLMs and fail to capture the complexity of a visual scene.

In this work, we investigate the ability of VLMs to handle tasks that involve complex visual reasoning by drawing inspiration from cognitive research on human reasoning [36]. Humans naturally engage in visual reasoning to navigate and interact with their surroundings, leveraging core cognitive abilities such as perception, spatial reasoning, and decision-making. Inspired by these mechanisms, we implement *visual chain-of-thought* as a structured rea-

*Equal contribution, with order determined alphabetically.

soning process that involves the following steps: detecting relevant entities in the scene, identifying their attributes, modeling spatial relationships, and reasoning to determine appropriate actions.

Existing reasoning-oriented visual question answering (VQA) datasets are insufficient for evaluating these capabilities. Some rely on synthetic or schematic visuals [5], lack grounding annotations [28, 29], or present oversimplified scenarios with a single region of interest [41]. Additionally, most explanations are generated from a set of predefined templates or by using an LLM [5, 30, 41]. These explanations can be repetitive, error-prone, and biased, limiting their effectiveness as training signals.

To overcome these limitations, we introduce DRIVINGVQA, a visual reasoning dataset derived from publicly available French driving theory exams. It contains 3,931 samples, each featuring one or two visual questions with multiple answer choices, alongside expert-written explanations and human annotations of relevant entities with bounding box coordinates. DRIVINGVQA offers a comprehensive framework to guide and assess VLMs' capabilities in complex real-world scenarios that involve multi-object and spatial reasoning.

Using DRIVINGVQA, we conduct a structured analysis to investigate the impact of various prompting and training strategies on VLM's reasoning capabilities, and propose *RIV-CoT*: *Retrieval-based Interleaved Visual Chain-of-Thought prompting* (see Fig. 1). While state-of-the-art VLMs struggle in zero-shot settings, we show that training strategies incorporating relevant entities information, in particular through visual crops interleaved with textual reasoning, significantly improve the correctness of reasoning and answer compared to direct answer generation and vanilla CoT prompting.

Our main contributions are as follows:

- We introduce DRIVINGVQA, a real-world VQA dataset designed to benchmark and enhance the visual reasoning capabilities of VLMs (Sec. 3).
- We propose *RIV-CoT*, a retrieval-based interleaved visual chain-of-thought method that allows VLMs to reason using visual crops retrieved from the input image (Sec. 4);
- We show that incorporating relevant entities information improves answer accuracy by up to 3.1% and reasoning accuracy by up to 4.6% compared to vanilla chain-of-thought prompting (Sec. 5.3).
- We demonstrate that our approach scales effectively to a larger dataset that does not include relevant entities annotation, A-OKVQA [40], by using automatically-generated pseudo-labels (Sec. 5.5).

2. Related Work

2.1. Visual Chain-of-Thought in VLMs

A major challenge in visual CoT for VLMs is their capacity to leverage visual inputs during the reasoning process. Significant research efforts have been made to enhance visual CoT capabilities. One of the main strategies is to generate a textual description of the scene before answering the question [20, 35, 39, 57]. The description may take the form of captions, a series of visual question-answer pairs that extract additional image details [58], or sparser representations of the input image such as scene graphs [32]. Similarly, visual programming [10, 19, 45] adopts a neuro-symbolic approach that leverages off-the-shelf models to extract information from images and convert it into text for subsequent reasoning.

Another research direction focuses on explicit grounding within the input image. This includes training VLMs to generate CoT along with coordinates of relevant image regions [4, 27], or incorporating visual prompts directly into the image, to be used by the VLM when reasoning. These visual prompts can take the form of a grid [16] or a set of visual markers [53] that indicate specific areas of the image. More closely related to our work, CogCOM [37] modifies the input image before using the transformed version to answer a question, notably by zooming on a specific image region. Concurrently, Visual CoT [41] designs a visual sampler that selects a sub-region of the input image to answer a question. In contrast, our method teaches VLMs to perform interleaved CoT with image patches retrieved from the input image, ultimately enabling more integrated visual reasoning.

2.2. Visual Reasoning Datasets

Since the seminal work of Antol et al. [2], visual question answering (VQA) datasets [7, 11, 31, 43] have flourished over the past decade. With the advent of VLMs, recent benchmarks [5, 29, 41, 54] are designed to evaluate deeper reasoning capabilities, particularly across multiple disciplines with domain-specific knowledge. Among datasets that provide explanations in addition to the question-answer (QA) pair, many comprise synthetic images with human-annotated explanations or real-world images paired with synthetically generated explanations. For instance, ScienceQA [29] offers human-annotated explanations sourced from elementary and high school science exams, but relies on figures, puzzles, or synthetic visuals. Similarly, PuzzleQA [5] is generated from a finite set of predefined templates and comprises synthetic visual puzzles, questions as well as explanations. Causal-VidQA [18] presents video-based questions accompanied by human-written explanation that require causal reasoning to be answered. However, it lacks the extensive entity annotations needed to ground

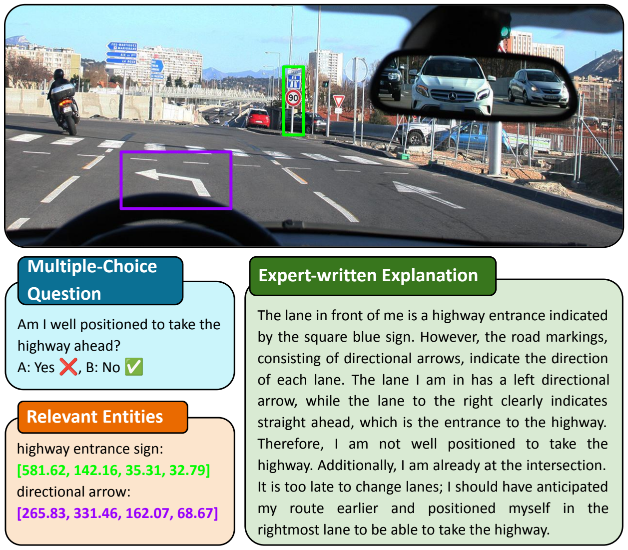
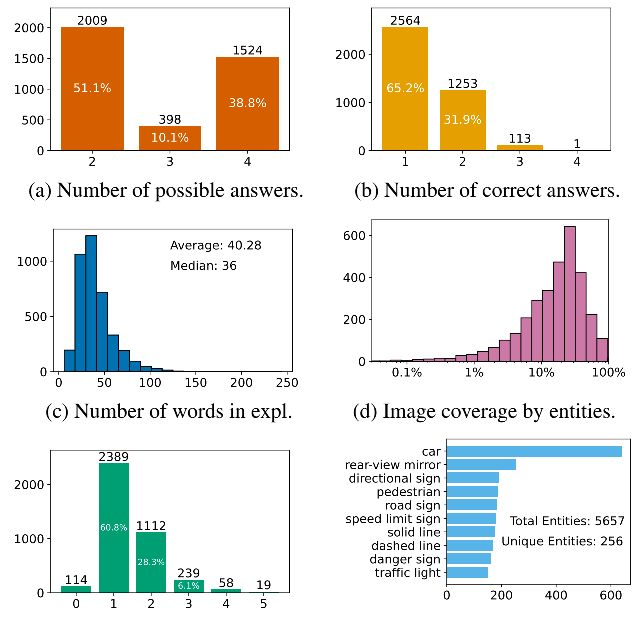


Figure 2. **DrivingVQA example** with a multiple-choice question, a set of relevant entities with their coordinates, and an expert-written explanation describing the situation step by step.



(a) Number of possible answers. (b) Number of correct answers.

(c) Number of words in expl. (d) Image coverage by entities.

(e) Number of relevant entities. (f) Top-10 most frequent entities.

Figure 3. **DrivingVQA dataset statistics.**

reasoning in visual elements. GQA-CoT [41] contains real-world images paired with their corresponding relevant entities but relies on synthetic explanations and restricts its annotations to a single entity per image.

In the field of autonomous driving, the need for explainable driving behavior in control planning [49, 52] and end-to-end driving [12, 50] has led to the creation of many VQA datasets [15, 30, 38] building upon existing autonomous driving datasets [38, 42]. NuScenesQA [38] provides simple, single-word language responses per question which hinders the possibility of complex reasoning. The recent LingoQA benchmark [30] comes closest to our work, featuring QA pairs annotated on an in-house dataset with textual descriptions of driver actions, justifications, and road observations. However, its questions and answers are synthetically generated by GPT-4 [6] and do not include any bounding boxes or relevant entities to visually ground explanations. In contrast, DrivingVQA features real-world samples enriched with human-expert annotations of relevant entities to answer the question and explanations (see Appendix D for a detailed dataset comparison).

3. DrivingVQA Dataset

We construct DrivingVQA through a two-step process: (1) collecting and filtering challenging real-world driving scenarios from driving theory tests (see Sec. 3.1), and (2) annotating relevant entities in each sample (see Sec. 3.2).

3.1. Data Collection

In many countries, obtaining a driver's license requires passing two examinations: a theoretical test and a practical driving test. In France, the theoretical exam consists of 40 multiple-choice questions (MCQs) with images. Candidates must possess accurate knowledge of traffic laws, road signs, and safe driving practices and must carefully analyze the image provided to answer the question. Given this emphasis on visual analysis and the use of real-world images, we collect French driving theory material.

Collection. To help candidates prepare for this exam, several online platforms offer practice driving theory problems. These resources are often freely accessible and do not require registration. We curate MCQs from such platforms. Each collected sample includes questions with two to four possible answers, where multiple answers may be correct. Some questions include two sub-questions, each with two possible answers. We standardize the format to include the image of the driving scene, the question's text, the list of possible answers, the correct answer(s), and an expert-annotated explanation. An example of a driving theory problem is shown in Fig. 2.

Filtering. Not all collected questions require visual reasoning about driving scenes, and/or some images may serve only illustrative purposes. We use GPT-4o to filter out samples where questions could be answered correctly without using the image, and perform a manual review of excluded images to address potential misclassifications. In a valida-

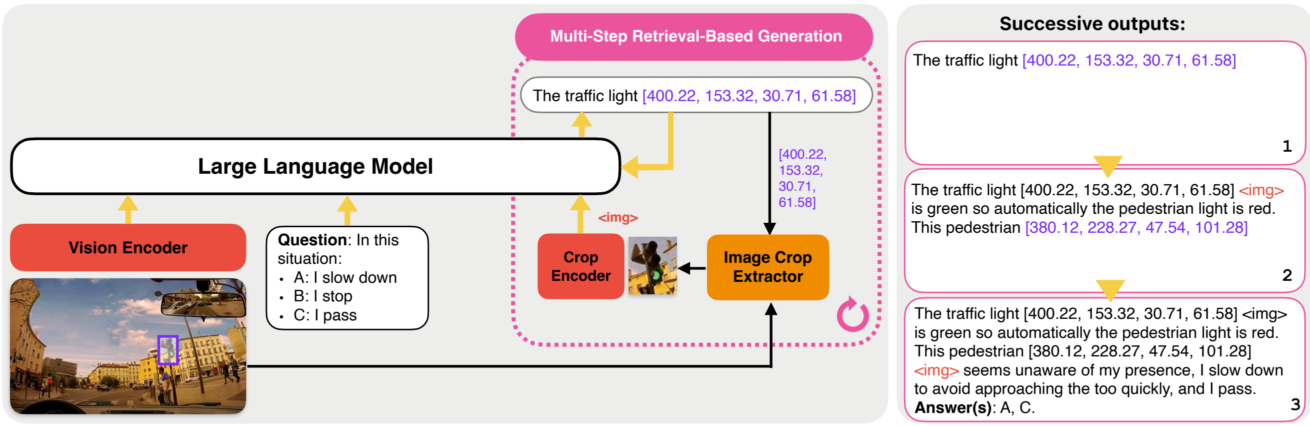


Figure 4. **Illustration of multi-step retrieval-based generation.** During inference, starting with a tokenized question and an image tokenized by the adapter on the output of the Vision Encoder, the Large Language Model generates output until it predicts a bounding box. At this point, the generation process pauses to extract the corresponding image crop based on the predicted coordinates. The image crop is encoded and adapted into an image crop token, which is then added back into the model's context along with the question, image tokens and previously generated outputs. This iterative process continues until the model produces its final answer.

tion exercise with 60 randomly selected samples, three expert annotators achieved an inter-annotator agreement score of 0.95 (Krippendorff's alpha, Fleiss' kappa, and Cohen's kappa).

Translation. We translate all questions, answer options, and explanations from French to English using GPT-4o-mini, then manually review and refine them for consistency and fidelity to the original intent.

After filtering, the final DRIVINGVQA dataset contains 3,931 samples, divided into a training set (80%) and a test set (20%, 789 samples). Fig. 3 provides an overview of the dataset statistics, including the distribution of the number of possible answers, correct answers, and explanation length in terms of word count.

3.2. Relevant Entities

Here, we recall our formalization of visual CoT: detecting relevant entities in the scene, identifying their attributes, modeling spatial relationships, and reasoning to determine appropriate actions. To support this, We augment DRIVINGVQA with annotations of relevant entities to answer the question, specifying their names and locations in the image.

Automated extraction. To reduce the burden of manual annotation, we develop a pipeline that identifies possible relevant entities and their location in the image. This process begins by leveraging human explanations to extract an initial list of domain-specific key entities found in our dataset. For each sample, we identify the entities from this list that are visible in the images and referenced in the question, possible answers, and explanation using GPT4o-mini. Then, we use GroundingDINO [25] to localize these entities in the image, obtaining (entity label, bounding box coordinates) pairs. Finally, we apply heuristics to refine these outputs, such as grouping similar labels under unified entity names. Details of this pipeline are provided in Appendix A.2.

Manual annotation. Human experts refine the pseudo-annotated data by removing irrelevant entities, correcting inaccurate labels and coordinates, and adding missing entities. In total, 5,657 entities (spanning 256 unique labels) were annotated with precise bounding boxes, averaging 1.4 entities per image. Fig. 3 shows the distribution of entities, their image coverage, and the top 10 most frequent entity labels.

4. Retrieval-Based Interleaved Visual CoT

In this section, we introduce *RIV-CoT* (Retrieval-based Interleaved Visual Chain-of-Thought), a framework designed to enhance the visual reasoning abilities of VLMs by explicitly grounding their chain-of-thought process in visual evidence. RIV-CoT relies on two distinct modules: (1) the data augmentation of human-annotated explanations by interleaving them with relevant entities, and (2) a multi-step retrieval-based generation during inference to dynamically insert image crops containing relevant entities detected by the VLM into its generated chain-of-thought.

4.1. Interleaved Explanation Augmentation

Explanations in DRIVINGVQA do not systematically make explicit reference to the human-annotated relevant entities. To ground explanations with visual cues, we integrate the list of relevant entities into the explanations using GPT-4o with two demonstrations, creating *interleaved* explanations. Each set of bounding-box coordinates is embedded into the

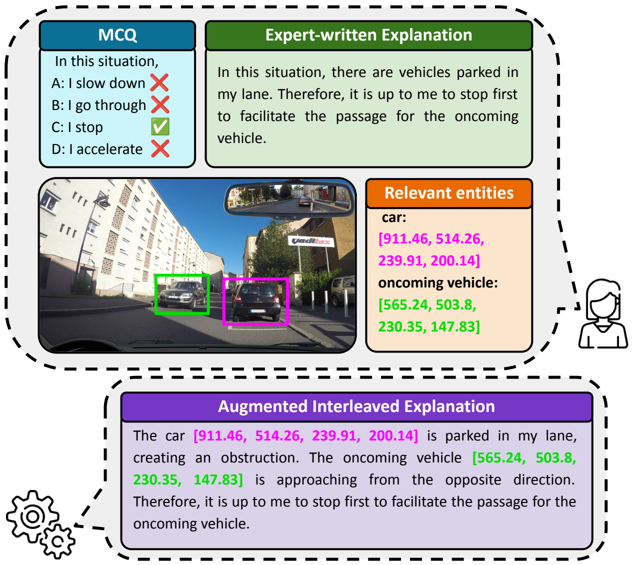


Figure 5. **Interleaved explanation augmentation.** We feed GPT-4o with the question and possible answers, the image, the list of relevant entities and coordinates, and the original expert-written explanation. The resulting interleaved explanation refers to the relevant entities early in the sentences, allowing the reasoning process to be conditioned on the content of the image crops.

explanation, reformulating it slightly to refer to the entity at the beginning of each step of the reasoning; thereby achieving a more causal formulation aligned with the autoregressive nature of the training objective (see example in Fig. 5). This process is followed by light rule-based post-processing to refine the final explanation (see details and prompt in Appendix A.1).

4.2. Multi-Step Retrieval-Based Generation

Leveraging the interleaved augmented explanations, we propose to train a VLM using a multi-turn conversation format where an explanation is broken down into sequential turns, with each model turn ending with the bounding box coordinates of a relevant entity. To capture additional context and better understand the entity's attributes, we expand the detected bounding box by 50%. The expanded image crop is extracted from the input image and given as input in the next turn, allowing the model to iteratively process visual and textual information in an interleaved fashion. Training is conducted using a standard autoregressive objective.

At inference, the generation process follows an iterative approach (see Fig. 4). The model generates outputs until it predicts a bounding box; at that point, the process pauses and the image crop corresponding to the predicted coordinates, expanded by 50%, is extracted. The cropped image is then encoded via the vision encoder and the patch tokens are inserted into the context after the bounding box, guiding the subsequent text generation. This multi-step retrieval and integration cycle continues until the final answer is produced, yielding intermediate steps that interleave image crops with textual reasoning, providing a more grounded CoT.

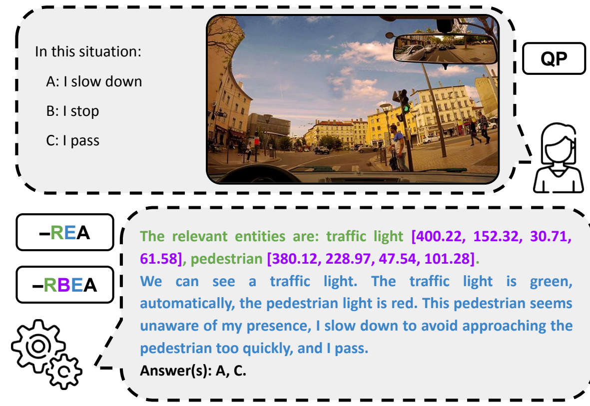


Figure 6. **Conversation formats for different fine-tuning strategies**. The model is given an image, a question and possible answer choices and is asked to predict relevant entities labels (**QP-REA**), or to predict labels and bounding box coordinates, before reasoning and answering (**QP-RBEA**).

5. Experiments

In this section, we first evaluate the zero-shot performance of state-of-the-art VLMs on DrivingVQA (Sec. 5.2). Then we evaluate RIV-CoT, exploring the benefits of incorporating entity-related information—such as entity names, spatial coordinates, and visual content—through supervised fine-tuning to enhance the model's reasoning abilities (Sec. 5.3). We complement these results with a fine-grained analysis of the quality of the generated entity coordinates and reasoning to gain insights into possible model limitations (Sec. 5.4). Finally, we extend our study to scenarios where high-quality entity annotations are unavailable, demonstrating how automatically generated pseudo-annotations can serve as an alternative for improving model performance at scale (Sec. 5.5).

5.1. Experimental Setup

With the exception of the zero-shot evaluation, all experiments involve fine-tuning the 7B version of LLaVA-OneVision (LLaVA-OV) [17] on the DrivingVQA train split. LLaVA-OV uses SigLIP [55] as image encoder, a two-layer MLP projector, and Qwen2 [48] as LLM backbone. All components are trained end-to-end with an autoregressive loss. Each training run is repeated five times with different random seeds to account for stochasticity in the fine-tuning process and all results are reported with standard deviation. Training is performed over 10

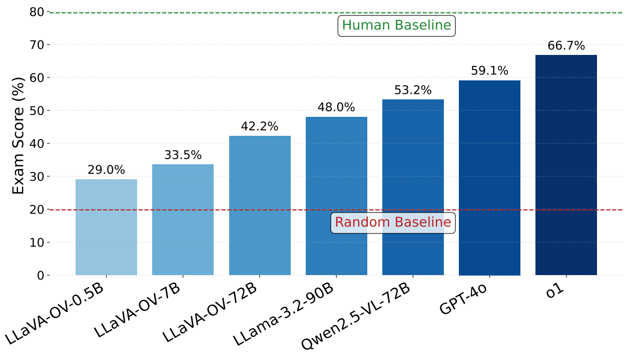


Figure 7. **Comparison of zero-shot model performance on the DRIVINGVQA test set**. The dashed lines indicate the random baseline (19.8%) and the human baseline (79.6%).

epochs, following hyperparameters used by the authors of LLaVA-OV [17].

Evaluation metrics. The performance on the DRIVINGVQA test split is measured with the *exam score*, analogous to the real driving theory score used to evaluate candidates: all correct answers must be selected to score a question correctly. Note that this score is equivalent to the exact match ratio. Since the task is analogous to multi-label classification, we also report the F1-Score.

Conversation format. To ease the understanding of what information is provided in the prompt and what the model generates, we use the following notation based on the components of DRIVINGVQA introduced in Sec. 3. All conversations begin with a prompt including at least the image, the question (**Q**) and the list of possible answers (**P**), and end with the predicted answer (**A**), potentially preceded by an explanation **(E)**. Hence, fine-tuning to directly answer the question is denoted as **DirectAnswer** (QP−A) while reasoning with the explanation before answering corresponds to **CoT** (QP−EA), where the dash (−) separates the input from the output of the VLM. Additional elements may include the list of the relevant entity labels (**R**), their bounding box coordinates (**B**), or the associated visual crops (**V**). The usage of interleaved explanation is denoted by **I**. Illustrative examples of conversation formats with relevant entities can be found in Fig. 6, while Fig. 5's augmented interleaved explanation showcases the IBEA format. Given this conversation format, **RIV-CoT** could also be denoted as QP−(IB−V)^N−EA, as it progressively incorporates multiple visual patches into the interleaved explanation across multiple steps.

Training Strategy	**Entity Format**				**Scores** (%)	
	Name	BBox	Vis.	Interlvd.	Exam	F1-Score
DirectAnswer					53.0 (±0.9)	63.3 (±0.6)
CoT					56.2 (±1.0)	65.8 (±0.9)
QP−REA	✓				57.0 (±1.3)	67.1 (±1.4)
QP−RBEA	✓	✓			57.7 (±0.7)	67.3 (±0.7)
QP−RB−RV−EA	✓	✓	✓		58.4 (±1.1)	67.8 (±1.1)
QP−IEA	✓	✓		✓	56.4 (±0.4)	66.3 (±0.6)
QP−IBEA	✓	✓		✓	57.9 (±0.5)	66.8 (±0.2)
RIV-CoT	✓	✓	✓	✓	**59.3** (±1.0)	**68.8** (±0.9)

Table 1. **Comparative results of fine-tuning LLaVA-OV-7B using different training strategies**. *Vis.* stands for "visual crops".

5.2. Zero-Shot Evaluation

In this section, we evaluate the zero-shot performance of popular VLMs on the DRIVINGVQA test set. For comparison, we include a random baseline where responses are selected randomly from all possible answer combinations for each question. Results are aggregated over 1,000 runs, reporting the mean exam and F1 scores. To estimate human performance, we recruit six participants with varying driving experience. They are asked to answer batches of 40 randomly selected samples in under 20 minutes, simulating the operational driving theory exam conditions. Their average exam score of 79.6% falls below the official passing score of 87.5% (35/40). This drop reflects the increased difficulty of DRIVINGVQA, as it focuses on visually challenging questions after the data filtering process.

Fig. 7 presents the zero-shot performance of various open-sourced models, including the LLaVA-OV variants (0.5B, 7B, 72B) used in the subsequent experiments, and proprietary models using DirectAnswer prompting. Larger models tend to perform better, with o1 [34] achieving the highest exam score (66.7%). Nevertheless, it remains far from the human baseline (79.6%), which illustrates the benchmark's difficulty due to its domain-specific images and knowledge and the complexity of the visual scenes. We also assess the impact of visual inputs in DRIVINGVQA by evaluating GPT-4o [33] without image information. This results in a substantial performance drop to 33.1% (-26 pts) compared to its image-enabled counterpart. It confirms that images are necessary to accurately answer DRIVINGVQA's questions, and knowledge-based shortcuts are not sufficient.

5.3. Fine-tuning with Relevant Entities

We explore how to enhance VLMs' visual chain-of-thought by leveraging relevant entities through their label, bounding box and visual crops, all of these interleaved in the explanation or not (Tab. 1); demonstrating the importance of each component of RIV-CoT.

First, we experiment with the **DirectAnswer** and chain-of-thought prompting (**CoT**) baselines. As expected, fine-tuning LLaVA-OV with **CoT** allows the model to generate better answers than predicting answers directly. While learning to generate the list of relevant entities (**QP−REA**) before generating an explanation and answer only leads to a minor improvement, adding bounding box coordinates (**QP−RBEA**) improves the exam score to 57.7% (+1.5 pts over **CoT**). Then, to leverage the visual content, we use a two-step reasoning format (**QP−RB−RV−EA**), where the model first predicts relevant entities' labels and coordinates given the question and answer choices (QP−RB). Then, it is fed with the predicted entities' visual patches, and uses it to reason and answer (−RV−EA).[1] This strategy further improves the performance, reaching 58.4% and showing the importance of visual patches for reasoning.

Finally, using explanations interleaved with visual patches, **RIV-CoT** achieves the best exam score (59.3%). This result shows that providing rich contextual visual information interleaved within explanations results in the most efficient way to enhance VLMs' reasoning abilities. Fine-tuning with interleaved explanation without coordinates nor visual patches (**QP−IEA**) leads to a score comparable to fine-tuning with the original explanations (**CoT**), validating that the performance gain seen in **QP−IBEA** and **RIV-CoT** comes from the added information from the bounding boxes and visual patches.

5.4. Analysis

We analyze the impact of the correctness of relevant entities' predicted coordinates on model performance, and the link between model performance, reasoning correctness and fine-tuning strategy.

5.4.1. Impact of entity detection correctness

The detection performance of models fine-tuned to predict bounding boxes of relevant entities is presented in Tab. 2. We measure the top-1 accuracy based on Hungarian matching at an IoU threshold of 0.50. Results are reported for all samples, as well as for samples with correct and incorrect final predictions.

For correctly answered samples, the detection accuracy is slightly higher than the overall average; while samples with incorrect predictions exhibit lower accuracy. Notably, **RIV-CoT** achieves the highest accuracy on correctly predicted samples (72.4%) but its detection performance decreases to 66.3% on incorrect answers, indicating that detection failures may contribute to prediction errors.

To further assess the effectiveness of visual patches, we compare models using predicted entities (**QP−RB−RV−EA**) against oracle entities (**QPRV−EA**) in Tab. 3. The oracle entities improve performance significantly (+3.3 pts in exam score, +3.1 pts in F1-score), highlighting a limitation in grounding capabilities of the VLM chosen in our experiment.

[1] Note that the training strategy of **QP−RB−RV−EA** shares similarities with Visual CoT [41] but uses a different visual patch cropping strategy and makes use of an unlimited number of relevant entities per sample.

Training Strategy	Top-1 Acc. @ IoU0.50 (%)		
	All	Correct	Incorrect
QP−RBEA	68.7 (±1.4)	69.3 (±1.1)	**67.6** (±1.4)
QP−IBEA	68.5 (±1.8)	68.9 (±1.7)	67.3 (±1.6)
QP−RB−RV−EA	**69.8** (±1.5)	72.3 (±1.4)	66.1 (±2.7)
RIV-CoT	69.6 (±1.7)	**72.4** (±1.2)	66.3 (±1.8)

Table 2. **Detection performance of bounding box-predicting models.** Top-1 accuracy at an IoU threshold of 50% is reported for all samples, correctly predicted, and incorrectly predicted samples.

Visual Patch Type	Exam (%)	F1-Score (%)
Predicted Entities (QP−RB−RV−EA)	58.7 (±0.8)	67.9 (±1.1)
Image Split (QPR−EA†)	60.7 (±0.9)	69.8 (±0.6)
Oracle Entities (QPRV−EA)	**62.0** (±0.4)	**71.0** (±0.5)

Table 3. **Performance analysis by type of visual patches.** Results are shown for models using predicted entities, oracle entities, and a multi-scale image splitting strategy. † indicates a model trained with AnyRes-4 [22] visual patches.

To determine whether this improvement stems from relevant visual information rather than merely more visual tokens, we compare **QPRV−EA** with a multi-scale image patching strategy (**QPR−EA**†), called *AnyRes* [22]. It consists in splitting the image into N sub-patches and concatenating all their representations with the original image, and is often used to encode high-resolution images. As the maximum number of entities per sample in DrivingVQA is five, we use $N = 4$ for a fair comparison. Results in Tab. 3 show that **QPR−EA**† only achieve 60.7% exam score, -1.3 pts compared to **QPRV−EA**, which confirms that carefully selected visual patches provide a stronger advantage than generic image splitting.

5.4.2. Reasoning Correctness

Our evaluation metrics, the exam score and the F1-score, measure a model's ability to select the correct answer. However, our approach is designed to enhance the model's VQA performance by making its reasoning more grounded in the input image. While a high exam score indicates an improved answer selection, it does not necessarily reflect the correctness of the reasoning. A model may rely on heuristics or memorization, a limitation that the exam score alone does not fully capture. To address this, we assess the correctness of the model's reasoning by comparing it against

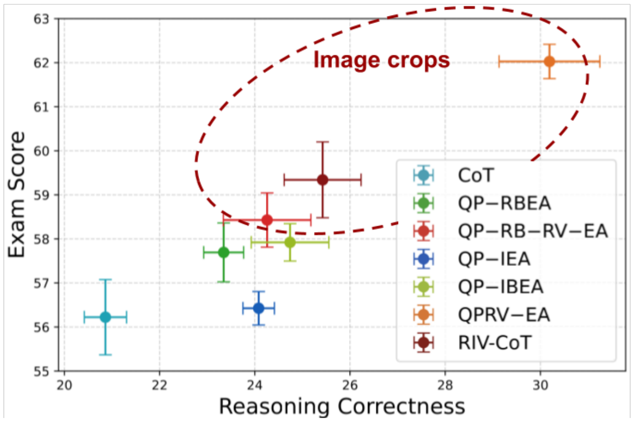


Figure 8. **Reasoning correctness vs. Exam score for each fine-tuning strategy.** In shades of green are methods using bounding boxes; in shades of red, using image crops.

DRIVINGVQA's ground truth explanations. We use GPT-4o-mini as an evaluator, leveraging the LLM-as-a-judge paradigm [59], which is increasingly used to scale the evaluation of open-ended generations of LLMs and VLMs, and is shown to align well with human judgment in pairwise comparisons [8, 26].

The prompt used for GPT-4o-mini is detailed in Appendix C.3. The model is provided with the question and answer choices, the ground truth explanation, and the reasoning generated by the VLM. It is then tasked with identifying the key arguments in both reasonings, checking for missing or contradictory elements, and determining whether they align. On a balanced set of 50 human-evaluated reasoning pairs sampled from different fine-tuning strategies, GPT-4o-mini achieves the same judging performance as GPT-4o, with an F1-score of 0.82. We define the *reasoning correctness* as the proportion of samples in the test set where the judge determines that the model's reasoning matches the ground truth reasoning.

We compare reasoning correctness and exam scores across our systems in Fig. 8. While the two metrics are positively correlated, the reasoning correctness consistently remains at least half of the exam score. The **CoT** baseline reaches a reasoning correctness of 20.8%; then come models incorporating bounding boxes (**QP−RBEA**, **QP−IBVEA**) and image crops (**QP−RB−RV−EA**, **RIV-CoT**). As a comparison, **QPRV−EA**, which receives image crops of oracle entities as input, achieves 30.2% reasoning correctness. Among retrieval-based approaches, **RIV-CoT** outperforms its non-interleaved counterpart **QP−RB−RV−EA** by 1.2 points and outperforms the **CoT** baseline by 4.6 points, demonstrating the benefits of visual patches and interleaved formatting for reasoning accuracy.

Dataset	Training Strategy	MC Acc. (%)
A-OKVQA	DirectAnswer	78.2 (±0.3)
	CoT	80.6 (±0.4)
	QP−RB−RV−CoT	82.3 (±0.3)
	RIV-CoT	**84.2** (±0.2)

Table 4. **Multiple-choice accuracy on the val subset of A-OKVQA [40]**, for different training strategies.

5.5. Scaling with Automatically-Extracted Entities

DRIVINGVQA includes high-quality annotated relevant entities along with their coordinates. Such annotations are rarely available, and transferring our method to new datasets requires external tools to obtain annotations. In this section, we extend our approach to A-OKVQA [40], a crowdsourced dataset composed of 25K visual questions requiring commonsense and world knowledge to be answered. Similar to DRIVINGVQA, each question in A-OKVQA is accompanied by multiple-choice options and rationales explaining the reasoning behind the correct answer. To obtain a list of relevant entities and their coordinates for each question, we employ an automated entity extraction pipeline based on the method introduced in Sec. 3.2. Specifically, we prompt GPT-4o-mini to generate potential relevant entity labels for each question. Then, we use the open-set object detector GroundingDINO [25] to localize these entities within the image. Only the top 5 entities with the highest confidence scores are retained.

Using these pseudo-labeled entities, we fine-tune LLaVA-OV-7B with our **RIV-CoT** method on the train set of A-OKVQA (17K samples), and evaluate its performance on the available validation subset (1992 samples). As shown in Tab. 4, **RIV-CoT** achieves a multiple-choice accuracy of 84.2%, outperforming both the **DirectAnswer** baseline (78.2%) and vanilla **CoT** prompting (80.6%). Notably, the two-step conversation strategy **QP−RB−RV−EA**, which integrates visual patches without interleaving them through retrieval-based generation, achieves 82.3% accuracy. These results highlight that our retrieval-based interleaved approach can largely improve the model's ability to accurately answer questions that require complex visual reasoning, even when relying on automatically extracted entities to compensate for the lack of gold-standard annotations.

6. Conclusion

In this work, we introduce DRIVINGVQA, a novel visual reasoning dataset derived from French driving theory exams, along with RIV-CoT, a retrieval-based interleaved visual chain-of-thought framework designed to enhance the visual reasoning abilities of VLMs. Our experiments on

DRIVINGVQA demonstrate that RIV-CoT significantly improves both answer accuracy and reasoning correctness compared to direct answer generation and vanilla chain-of-thought prompting. Furthermore, we show that our approach scales effectively to datasets lacking human annotations by leveraging automatically generated pseudo-labels, ensuring broader applicability in real-world scenarios.

Despite these promising advances, our analysis highlights challenges that VLMs face in zero-shot settings. A promising direction for future research is to integrate retrieval-based interleaved visual chain-of-thought reasoning into VLM's pretraining, by scaling the automatically extracted entities variant on extensive grounding datasets and established multimodal reasoning benchmarks.

Acknowledgments

We thank Max Luca Pio Conti, Pierre Ancey, Francesco Pettenon and Matthias Wyss for their contributions to preliminary work. We thank Auguste Poiroux, Gaston Lenczner, Florent Forest, Jacques Everwyn, Vincent Montariol, Alice Legrand, Marc Lafon, Yannis Karmim, and Alexandre Merkli for the human evaluation of the DRIVINGVQA test set. We also thank the VITA lab members for their valuable feedback, which helped to enhance the quality of this manuscript. SM gratefully acknowledges the support of the Swiss National Science Foundation (No. 224881). AB gratefully acknowledges the support of the Swiss National Science Foundation (No. 215390), Innosuisse (PFFS-21-29), the EPFL Center for Imaging, Sony Group Corporation, and the Allen Institute for AI.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 1

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015. 2

[3] Shivam Chandhok, Wan-Cyuan Fan, and Leonid Sigal. Response wide shut: Surprising observations in basic vision language model capabilities. *arXiv preprint arXiv:2408.06721*, 2024. 1

[4] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 1, 2, 7

[5] Yew Ken Chia, Vernon Toh Yan Han, Deepanway Ghosal, Lidong Bing, and Soujanya Poria. Puzzlevqa: Diagnosing multimodal reasoning challenges of language models with abstract visual patterns. *arXiv preprint arXiv:2403.13315*, 2024. 2, 13

[6] OpenAI Contributors. Gpt-4 technical report, 2024. 3

[7] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[8] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024. 8

[9] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14375–14385, 2024. 1

[10] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14953–14962, 2023. 2

[11] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[12] Jyh-Jing Hwang, Runsheng Xu, Hubert Lin, Wei-Chih Hung, Jingwei Ji, Kristy Choi, Di Huang, Tong He, Paul Covington, Benjamin Sapp, Yin Zhou, James Guo, Dragomir Anguelov, and Mingxing Tan. Emma: End-to-end multimodal model for autonomous driving. *arXiv preprint arXiv:2410.23262*, 2024. 3

[13] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. 1

[14] Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27036–27046, 2024. 1

[15] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for self-driving vehicles. In *European Conference on Computer Vision (ECCV)*, 2018. 3, 13

[16] Xuanyu Lei, Zonghan Yang, Xinrui Chen, Peng Li, and Yang Liu. Scaffolding coordinates to promote vision-language coordination in large multi-modal models. *arXiv preprint arXiv:2402.12058*, 2024. 1, 2

[17] 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. 5, 6

[18] Jiangtong Li, Li Niu, and Liqing Zhang. From representation to reasoning: Towards both evidence and commonsense reasoning for video question-answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 13

[19] Xue Li, Yiyou Sun, Wei Cheng, Yinglun Zhu, and Haifeng Chen. Chain-of-region: Visual language models need details for diagram analysis. In *The Thirteenth International Conference on Learning Representations*, 2025. 2

[20] Zhiyuan Li, Dongnan Liu, Chaoyi Zhang, Heng Wang, Tengfei Xue, and Weidong Cai. Enhancing advanced visual reasoning ability of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1915–1929, 2024. 1, 2

[21] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26689–26699, 2024. 1

[22] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 7

[23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 1

[24] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024. 1

[25] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 4, 8, 3

[26] Yinhong Liu, Han Zhou, Zhijiang Guo, Ehsan Shareghi, Ivan Vulić, Anna Korhonen, and Nigel Collier. Aligning with human judgement: The role of pairwise preference in large language model evaluators. *arXiv preprint arXiv:2403.16950*, 2024. 8

[27] Yuecheng Liu, Dafeng Chi, Shiguang Wu, Zhanguang Zhang, Yaochen Hu, Lingfeng Zhang, Yingxue Zhang, Shuang Wu, Tongtong Cao, Guowei Huang, et al. Spatialcot: Advancing spatial reasoning through coordinate alignment and chain-of-thought for embodied task planning. *arXiv preprint arXiv:2501.10074*, 2025. 1, 2

[28] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, pages 216–233. Springer, 2025. 2

[29] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 2, 13

[30] Ana-Maria Marcu, Long Chen, Jan Hünermann, Alice Karnsund, Benoit Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, Elahe Arani, and Oleg Sinavski. Lingoqa: Visual question answering for autonomous driving. In *European Conference on Computer Vision (ECCV)*, 2024. 2, 3, 13

[31] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[32] Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14420–14431, 2024. 1, 2

[33] OpenAI. Gpt-4o system card, 2024. 1, 6

[34] OpenAI. Openai o1 system card, 2024. 6

[35] Övgü Özdemir and Erdem Akagündüz. Enhancing visual question answering through question-driven image captions as prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1562–1571, 2024. 1, 2

[36] Jean Piaget. *Piaget's theory of intelligence*. Englewood Cliffs, NJ: Prentice Hall, 1978. 1

[37] Ji Qi, Ming Ding, Weihan Wang, Yushi Bai, Qingsong Lv, Wenyi Hong, Bin Xu, Lei Hou, Juanzi Li, Yuxiao Dong, et al. Cogcom: Train large vision-language models diving into details through chain of manipulations. *arXiv preprint arXiv:2402.04236*, 2024. 2

[38] Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario, 2024. 3, 13

[39] Yuxuan Qiao, Haodong Duan, Xinyu Fang, Junming Yang, Lin Chen, Songyang Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. Prism: A framework for decoupling and assessing the capabilities of vlms. *Advances in Neural Information Processing Systems*, 37:111863–111898, 2025. 2

[40] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision (ECCV)*, 2022. 2, 8

[41] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *arXiv preprint arXiv:2403.16999*, 2024. 1, 2, 3, 7, 13

[42] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. In *European Conference on Computer Vision (ECCV)*, 2024. 3, 13

[43] Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE*

Conference on Computer Vision and Pattern Recognition (CVPR), 2019. 2

[44] Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. *arXiv preprint arXiv:2409.12183*, 2024. 1

[45] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11888–11898, 2023. 2

[46] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1

[47] 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. 1

[48] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 5

[49] Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and Jose M. Alvarez. Omnidrive: A holistic llm-agent framework for autonomous driving with 3d perception, reasoning and planning, 2024. 3

[50] Tsun-Hsuan Wang, Alaa Maalouf, Wei Xiao, Yutong Ban, Alexander Amini, Guy Rosman, Sertac Karaman, and Daniela Rus. Drive anywhere: Generalizable end-to-end autonomous driving with multi-modal foundation models, 2023. 3

[51] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1

[52] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*, 2024. 3

[53] An Yan, Zhengyuan Yang, Junda Wu, Wanrong Zhu, Jianwei Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Julian McAuley, Jianfeng Gao, et al. List items one by one: A new data source and learning paradigm for multimodal llms. *arXiv preprint arXiv:2404.16375*, 2024. 2

[54] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[55] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023. 5

[56] Zhuosheng Zhang, Aston Zhang, Mu Li, hai zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research (TMLR)*, 2024. 1

[57] Zhuosheng Zhang, Aston Zhang, Mu Li, George Karypis, Alex Smola, et al. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*, 2024. 1, 2

[58] Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *Advances in Neural Information Processing Systems*, 36:5168–5191, 2023. 2

[59] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023. 8

Retrieval-Based Interleaved Visual Chain-of-Thought in Real-World Driving Scenarios

Supplementary Material

This supplementary material contains the following sections:

- A description of our pipeline to generate the interleaved explanations in our dataset (Appendix A.1);
- A description of our pipeline to annotate relevant entities (Appendix A.2);
- Examples showcasing representative samples and interleaved explanations (Appendix B);
- Detail of the prompts used for training our models (Appendix C.1);
- Implementation details and hyperparameter configurations for model training (Appendix C.2);
- LLM-as-a-judge prompt for evaluation of reasoning correctness (Appendix C.3);
- Examples comparing various model outputs, showcasing the strengths and limits of our methods (Appendix C.4);
- A table comparing DRIVINGVQA with visual reasoning and autonomous driving datasets from the literature (Appendix D).

A. Dataset Specifications

This section details the methodologies employed for the generation of interleaved explanations and the annotations of relevant entities.

A.1. Pipeline for Generating Interleaved Explanations

As described in Sec. 3.2, for each visual question in our dataset, human experts use the explanations to identify and localize the key entities in the image that are required for answering the question. This leads to an average of 1.5 entities per image, and up to 5 entities. Then, as explained in Sec. 4.1, we use GPT-4o to match this list of relevant entities back with the explanation to obtain an interleaved explanation. In practice, we want to interleave each key entity – whether it is the entity label, its bounding box coordinates, or the corresponding image tokens – inside the explanation. In the rest of the section, we represent the interleaved explanation with bounding box coordinates appended next to the name of the entity referred to in the explanation.

To generate the interleaved explanations, we employ a strategy that combines few-shot prompting and cleaning heuristics.

Initial interleaved explanation generation. First, we feed GPT4o with each sample: the image, question, list of options, explanation, and the list of manually annotated entities along with bounding box coordinates. Following two demonstrations and a strict set of instructions, it is asked to interleave the bounding boxes of the entities inside the explanation, with minimal change to the explanation. Here, we provide the full instructions and the two hand-crafted demonstrations fed to the model.

System prompt

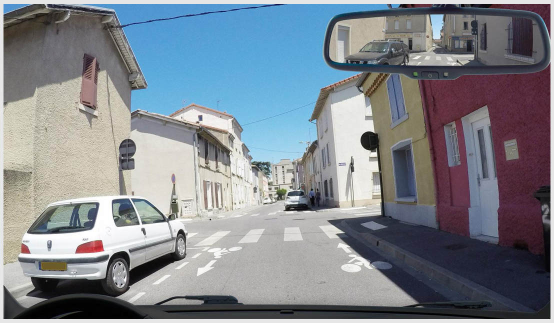
You are an expert at driving theory. You are tasked with helping a student answer questions about driving scenes.

Instructions

You are given an image of the driving theory test, a question about this image, a list of options, and the reasoning leading to the correct answer. I also give you a list of entities along with bounding box coordinates, that are relevant to answering the question. I need you to:

1. Convert the original reasoning into a clear, step-by-step reasoning that makes use of all entities in the list.
2. If the original reasoning does not mention one of the entities at all, add a short sentence referencing that entity and link it with the reasoning steps.
3. Refer to each relevant entity in the format: **entity_name** [x1, x2, y1, y2]. Replicate bounding box coordinates exactly as provided in the list.
4. Keep all other wording as close to the original reasoning as possible.
5. The entities should always be mentioned at the beginning of the sentences.

Demonstration 1



Question: ["I am turning right at the next intersection. My attention is focused on: pedestrians", "I am turning right at the next intersection. My attention is focused on: cyclists"]
Options: {A: Yes, B: No, C: Yes, D: No}
Answers: A, C
Reasoning: Before turning right, I will approach a crosswalk. My attention is therefore on the arrival of pedestrians. As I am alongside a bike lane, I will check my right blind spot before turning my wheels to ensure that no cyclist is next to me.
Entities:

- **pedestrian crossing**: [457.51, 595.61, 499.56, 62.67]
- **bike lane marking**: [893.4, 597.64, 196.2, 202.08]

Interleaved Reasoning

Before turning right at the next intersection, I will approach a **pedestrian crossing** [359.5, 337.11, 259.48, 22.09]. My attention is therefore on the arrival of pedestrians. As I am alongside a **bike lane marking** [893.4, 597.64, 196.2, 202.08], I will check my right blind spot before turning my wheels to ensure that no cyclist is next to me.

Demonstration 2

Question: By taking this exit
Options: {A: I can slow down right now, B: I must wait to brake}
Answers: A
Reasoning: A deceleration lane allows me to exit without disrupting the flow of traffic behind me. The vehicle behind is far away, so I can start slowing down right now.
Entities:

- **deceleration lane:** [933.62, 522.54, 537.35, 249.17]
- **exit sign:** [904.6, 413.47, 57.03, 36.02]
- **vehicle:** [1206.9, 156.54, 44.8, 28.56]
- **rear-view mirror:** [952.52, 82.74, 544.27, 181.17]

Interleaved Reasoning

A **deceleration lane** [933.62, 522.54, 537.35, 249.17] allows me to exit without disrupting the flow of traffic behind me. An **exit sign** [904.6, 413.47, 57.03, 36.02] indicates the upcoming exit. The **vehicle** [1206.9, 156.54, 44.8, 28.56] behind is far away, as I can see in the **rear-view mirror** [952.52, 82.74, 544.27, 181.17]. So I can start slowing down right now.

Explanations filtering and cleaning. We clean the generated interleaved explanations using regular expressions and heuristics.

- We match the bounding box coordinates in the generated interleaved explanation with the ones in the list of entities provided, correcting minor variations due to the model failing to exactly replicate the set of coordinates.
- We remove any hallucinated set of coordinates, that is absent from the provided list of entities.
- When an annotated bounding box was used twice in the interleaved explanation with different entity labels; if we can automatically identify the correct entity label, we remove the duplicated bounding box coordinates. Otherwise, we keep only the first occurrence of the set of coordinates.

A.2. Pipeline for Annotating Relevant Entities

To enrich the collected dataset with relevant entities for each sample, we developed a semi-automated entity extraction and localization pipeline that helped to accelerate the annotation process. This pipeline comprises three steps, detailed below.

Initial entity extraction. We define a taxonomy of potential entities commonly encountered in driving scenarios, organized into six groups (see Tab. 5). This taxonomy includes categories such as road signs, road markings, vehicles, people and other objects. Given this taxonomy, a multi-modal language model, such as GPT-4o-mini, is prompted with instructions to extract for each sample an initial list of entities relevant to the human-expert explanation. The prompt also integrates textual cues from questions, possible answers, and correct answer along with the associated image. The prompt, illustrated in Fig. 9, is designed to guide the model to prioritize visible and contextually significant entities. The extracted entities are returned in a structured list format, e.g., [`cyclist, pedestrian crossing, solid line`].

Relevant entities localization. After generating the list of relevant entities, we use a pre-trained object detection model, such as GroundingDINO [25], to localize these entities within the images. This step provides a bounding box for each detected entity. Detected entities undergo refinement such as grouping semantically similar labels under a unified category and filtering out irrelevant or erroneous detections that deviate from the predefined taxonomy.

Manual validation. The output of the automated pipeline is validated and refined by human annotators. Irrelevant entities are removed, inaccurate labels are corrected, and missing entities are added to ensure dataset consistency and accuracy.

Category	Entities
Road Signs	speed limit sign, end of speed limit sign, yield sign, directional sign, stop sign, intersection sign, mandatory right turn sign, mandatory left turn sign, mandatory straight ahead sign, no entry sign, no right turn sign, no left turn sign, no u-turn sign, no overtaking sign, end of overtaking prohibition sign, danger sign, priority sign, exit sign, dead end road sign, merge lane sign, level crossing sign, two-way traffic sign, emergency phone sign, handicapped accessible sign, parking prohibition sign, end of restrictions sign, dimension restriction sign, road narrowing sign, one-way street sign, construction detour sign, pedestrian crossing sign, pedestrian underpass sign, school crossing sign, town entry sign, town exit sign, direction sign, wild animal crossing sign, construction sign, toll road sign, weather-related sign, camping zone sign, chevron alignment marker
Road Markings	solid line, dashed line, pedestrian crossing, directional arrow, merge lane arrow, bike lane marking, stop line marking, loading zone line marking, traffic cones, temporary barrier
Road Features	speed bump, roundabout, tunnel, bridge, construction zone, accident, emergency phone, toll booth, parking lane, bus lane, bus stop area, bicycle lane, emergency lane, entry lane, exit lane
Vehicles	car, truck, motorcycle, bus, cyclist, van, motorhome, agricultural vehicle, public service vehicle, emergency vehicle
People and Animals	pedestrian, police officer, construction worker, horse rider, animal
Vehicle Parts	rear-view mirror, side-view mirror, turn signals, brake lights

Table 5. Categorized taxonomy of relevant entities for driving scenarios.

You are a driving theory expert, and your role is to extract entities from a driving scenario. These entities will be passed to an object detector for recognition.

All the possible entities are:
- road signs: (See listed entities in Tab. 5)
- road markings: (See listed entities in Tab. 5)
- road features: (See listed entities in Tab. 5)
- vehicles: (See listed entities in Tab. 5)
- people and animals: (See listed entities in Tab. 5)
- vehicle parts: (See listed entities in Tab. 5)

Instructions
Extract all entities from the scenes that are relevant to the following explanation and return them as a list. The output format should be only a list of entities, such as `[cyclist, oncoming vehicle, solid line, pedestrian crossing]`. Prioritize visible signs, markings, and vehicles directly affecting the scenario. If present in the image, always include any rear-view mirror or side-view mirror. For help, you can also refer to the questions, possible answers and true answer below, as well as the provided image attached.

Question(s): <questions_text>
Possible answers: <answers_text>
Correct answer: <correct_answer_letters>
Explanation: <explanation_text>

Figure 9. Prompt for relevant entity extraction.

B. Examples

This section presents four representative examples from DRIVINGVQA dataset. Each example includes an egocentric image, one or two questions, a set of 2 to 4 answer options, the correct answers, a list of entities critical for answering the questions, and a reasoning explanation interleaved with the relevant entities. We also provide the original non-interleaved reasoning to enable a direct comparison.

The first two examples showcase where different entities have the same label, making the matching more challenging. Note that these two examples have question pairs, the first one associated with answer choices A and B, the second one having answer choices C and D.

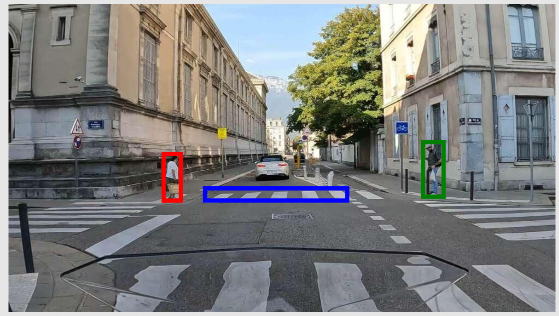
Augmented Dataset Example 1

Question: ["Can I drive at 50 km/h:", "70 km/h"]
Options: {A: Yes, B: No, C: Yes, D: No}
Answers: A, C
Entities:

- **speed limit sign**: [429.54, 50.63, 38.06, 35.67]
- **speed limit sign**: [431.53, 86.7, 33.9, 49.8]

Reasoning: The 2 speed limit signs are independent. The lower one limits the speed to 50 km/h only for vehicles designated for the transport of goods, as indicated by the category sign. The 70 km/h speed limit sign applies to all other categories of vehicles. I am not driving a goods transport vehicle, so I can drive at any speed not exceeding 70 km/h.
Interleaved Reasoning: The two speed limit signs are independent. The lower **speed limit sign** [429.54, 50.63, 38.06, 35.67] limits the speed to 50 km/h only for vehicles designated for the transport of goods, as indicated by the category sign. The **speed limit sign** [431.53, 86.7, 33.9, 49.8] applies to all other categories of vehicles. I am not driving a goods transport vehicle, so I can drive at any speed not exceeding 70 km/h.

Augmented Dataset Example 2



Question: ["I monitor the pedestrian's intention:", "I prepare to stop:"]
Options: {A: from the left, B: from the right, C: Yes, D: No}
Answers: A, B, C
Entities:

- **pedestrian**: [284.17, 274.3, 31.37, 84.98]
- **pedestrian crossing**: [359.5, 337.11, 259.48, 22.09]
- **pedestrian**: [757.46, 252.83, 37.6, 99.3]

Reasoning: The pedestrian on the left is very close to the crosswalk and is therefore preparing to cross. The pedestrian on the right is leaning over his phone and his attention is significantly diminished, so I am also monitoring him. I prepare to stop to let these two pedestrians cross.
Interleaved Reasoning: The **pedestrian** [284.17, 274.3, 31.37, 84.98] on the left is very close to the **pedestrian crossing** [359.5, 337.11, 259.48, 22.09] and is therefore preparing to cross. The **pedestrian** [757.46, 252.83, 37.6, 99.3] on the right is leaning over his phone, and his attention is significantly diminished, so I am also monitoring him. I prepare to stop to let these two pedestrians cross.

The following example shows a case where the explanation had to be modified, adding an extra sentence to include the relevant entity *dashed line* that was manually annotated by the human experts but wasn't mentioned in the original explanation.

Augemented Dataset Example 3

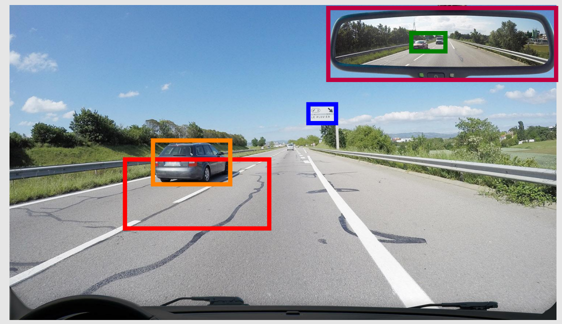
Question: Do the tradespeople run a risk if they park their van in the same way these police vans are parked?
Options: {A: Yes, B: No}
Answer: A
Entities:

- **police vehicle**: [372.0, 235.7, 1562.24, 1043.57]
- **bus lane**: [14.46, 755.5, 393.74, 996.79]
- **dashed line**: [320.34, 747.66, 214.84, 998.63]

Reasoning: These vans are not parked properly: they overflow so much from the spaces that road users have to move onto the bus lane, which is reserved for buses. We advise an average tradesperson against parking their van in this manner.
Interleaved Reasoning: These **police vehicles** [372.0, 235.7, 1562.24, 1043.57] are not parked properly: they overflow so much from the spaces that road users have to move onto the **bus lane** [14.46, 755.5, 393.74, 996.79], which is reserved for buses. The **dashed line** [320.34, 747.66, 214.84, 998.63] indicates the boundary of the lane. We advise an average tradesperson against parking their van in this manner.

This final example shows a case with 5 annotated relevant entities, which is the maximal number that can be found in our dataset.

Augmented Dataset Example 4



Question: ["The vehicle in front can still take the next exit", "I can follow this vehicle to overtake"]
Options: {A: Yes, B: No, C: Yes, D: No}
Answers: B, D
Entities:

- **dashed line**: [317.36, 423.1, 394.94, 188.64]
- **exit sign**: [819.58, 272.31, 75.71, 51.58]
- **car**: [1101.09, 77.8, 93.61, 49.79]
- **rear-view mirror**: [873.99, 9.21, 623.74, 196.48]
- **car**: [394.01, 371.19, 211.56, 119.92]

Reasoning: The marking in the middle of the two lanes of traffic only prohibits vehicles in the left lane from merging to take the exit. However, vehicles are approaching from behind, so I must give up overtaking.
Interleaved Reasoning: The **dashed line** [317.36, 423.1, 394.94, 188.64] in the middle of the two lanes of traffic only prohibits **vehicles in the left lane** [394.01, 371.19, 211.56, 119.92] from merging to take the exit indicated by the **exit sign** [819.58, 272.31, 75.71, 51.58]. However, vehicles [**1101.09, 77.8, 93.61, 49.79**] are approaching from behind, as seen in the **rear-view mirror** [873.99, 9.21, 623.74, 196.48], so I must give up overtaking.

C. Experiments Details

C.1. Format-Specific Prompt Templates

We begin by providing the detailed prompting format for each variant (QPR-EA, QPRB-EA, QPRV-EA) to analyze which signal is the most relevant to enhance visual-chain-of-thought reasoning. Note that the bounding boxes are represented in the format (x, y, w, h), where (x, y) specifies the coordinates of the top-left corner, and w and h denote the width and height of the bounding box. Similarly to [4] we normalize the bounding boxes between 0 and 1 during training and generation. The following example illustrates the QPR-EA and the QPRB-EA prompting strategies:

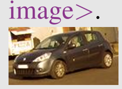
QPRB-EA

Human:
Unless explicitly stated otherwise, assume you are driving a car in France.
Select all correct answers to the following question from the available options. Detail your reasoning step by step based on road signs, markings, signals, and relevant driving rules. Then, provide the letters corresponding to your answer in the format: ’Answer(s): <letters>’
Question: As I approach this intersection, I position my foot above the brake Options: {A: Yes, B No.}
The relevant entities for this problem are: car [0.715, 0.665, 0.908, 0.830].
VLM:
Reasoning: At the level of this intersection, there is a car on the right. I don’t know if the driver has seen me since he is not looking at me. Therefore, I position my foot in a safe position above the brake so that I can stop if the driver decides to proceed.
Answer(s): A.

For QPRV-EA, we use a prompting strategy with visual tokens denoted <image> that is the features of the relevant entity’s bounding box from the SigLIP visual encoder. It corresponds to:

QPRV-EA

Human:
Unless explicitly stated otherwise, assume you are driving a car in France.
Select all correct answers to the following question from the available options. Detail your reasoning step by step based on road signs, markings, signals, and relevant driving rules. Then, provide the letters corresponding to your answer in the format: ’Answer(s): <letters>’
Question: As I approach this intersection, I position my foot above the brake:
Options: {A: Yes, B No.}
The relevant entities for this problem are: car < image>.



VLM:
Reasoning: At the level of this intersection, there is a car on the right. I don’t know if the driver has seen me since he is not looking at me. Therefore, I position my foot in a safe position above the brake so that I can stop if the driver decides to proceed.
Answer(s): A.

Then, we provide the prompts used for the two-step procedure outlined in Sec. 5.3 to evaluate the ability of VLMs to predict relevant entities under different formats (name: QP-REA, localization: QP-RBEA, visual content: QP-RB-RV-EA). The following illustrates QR-REA and QP-RBEA:

QP-RBEA

Human:
Unless explicitly stated otherwise, assume you are driving a car in France. List all relevant entities from the scene that are necessary to answer the following question, such as road signs, markings, signals, or other vehicles in the image, along with their bounding boxes. Then, select all correct answers to the following question from the available options. Detail your reasoning step by step based on these entities and relevant driving rules. Provide the letters corresponding to your answer in the format: 'Answer(s): <letters>'
Question: As I approach this intersection, I position my foot above the brake:
Options: {A: Yes, B No.}
VLM:
The relevant entities for this problem are: car [0.715, 0.665, 0.908, 0.830].
Reasoning: At the level of this intersection, there is a car on the right. I don't know if the driver has seen me since he is not looking at me. Therefore, I position my foot in a safe position above the brake so that I can stop if the driver decides to proceed.
Answer(s): A.

Finally, the example below shows how to prompt the two-step procedure for QP-RB-RV-EA.

QP-RB-RV-EA

Human:
Unless explicitly stated otherwise, assume you are driving a car in France.
List all relevant entities from the scene that are necessary to answer the following questions, such as road signs, markings, signals, or other vehicles in the image
Question: As I approach this intersection, I position my foot above the brake:
Options: {A: Yes, B No.}
VLM:
The relevant entities for this problem are: car [0.715, 0.665, 0.908, 0.830].
Human:
Their corresponding image patches are: car <image>. 
Then, select all correct answers to the following questions from the available options. choose at least one answer per question. Detail your reasoning step by step based on these entities and relevant driving rules. Provide the letters corresponding to your answer in the format: 'Answer(s): <letters>'.
VLM:
Reasoning: At the level of this intersection, there is a car on the right. I don't know if the driver has seen me since he is not looking at me. Therefore, I position my foot in a safe position above the brake so that I can stop if the driver decides to proceed.
Answer(s): A.

C.2. Implementation Details

We fine-tune various LLaVA-OneVision models using different prompt formats on the DRIVINGVQA train set. All models are initialized from the pre-trained weights obtained after the final OneVision training stage.

The fine-tuning experiments are conducted using 2 NVIDIA A100 GPUs, except for experiments involving visual patches ('V') where 4 NVIDIA A100 GPUs were required. We employ the following configuration and hyperparameters:

Hyperparameter	Value
Epochs	10
Batch Size	2
Max Seq. Length	32768
Vision Tower	SigLIP-SO400M-Patch14-384
Language Model	Qwen2-7B
Projector	MLP2x_GELU
Trainable Parts	ϕ_{vision}, θ_{proj}, θ_{LLM}
Learning Rate θ_{proj}, θ_{LLM}	$1e^{-5}$
Learning Rate ϕ_{vision}	$2e^{-6}$
Weight Decay	0
LR Scheduler	cosine
Warm-Up Ratio	0.03

Table 6. Hyper-parameter configuration for fine-tuning LLaVA-OneVision on the DRIVINGVQA dataset

C.3. Model reasoning evaluation

C.4. Result Output Examples

In this section, we present test samples alongside selected model outputs to highlight their capabilities and their limitations. For each sample, we provide the question and answer choices based on the QP-EA prompt (blue box), followed by the reasoning and answers from two models, along with their prompt-specific formats where applicable (red boxes).

Question Example 1

Human:
Unless explicitly stated otherwise, assume you are driving a car in France.
Select all correct answers to the following question from the available options. Detail your reasoning step by step based on road signs, markings, signals, and relevant driving rules. Then, provide the letters corresponding to your answer in the format: 'Answer(s): <letters>'
Question: To go straight, can I stay in this lane?
Options: {A: Yes, B No.}

Question Ouput 1: **QP-EA**

QP-EA:
Reasoning: If I want to go straight, the road markings indicate that I am correctly positioned, as I can either continue straight or turn right. The left lane is the lane for turning left. Therefore, I will not need to change lanes.
Answer(s): A.

Question Ouput 1: **QPRV-EA**

Human:
The relevant entities for this problem are: directional arrow .

QPRV-EA:
Reasoning: The lane I am in is intended for users wishing to turn right. This is indicated by the directional arrow on the road markings. Therefore, I cannot stay in this lane if my intention is to go straight.
Answer(s): B.

In Example 1 above, the simple chain-of-thought model (QP-EA) acknowledges the existence of a directional arrow as road marking, but incorrectly hallucinates its direction, leading to the wrong answer. The "Oracle" model QPRV-EA, being fed with an image patch showing the arrow, is

Instructions

You are a strict but fair driving-theory instructor. You're given:
1. A driving theory test question
2. A list of possible answer options
3. The official "correct reasoning"
4. A student's reasoning for the same question

Your task: *Assess if the student's reasoning matches the correct reasoning.*

Step-by-Step Instructions:
1. **Identify Student's Arguments**
   - List each key argument or step in the student's reasoning.
   - For each argument, briefly state whether it is correct or not, given the provided correct reasoning.
2. **Check for Missing or Contradictory Points**
   - Look at the official correct reasoning.
   - List **important points or steps** from the correct reasoning that the student **omits** or **directly contradicts**.
   - Minor omissions or differences in wording/style are acceptable.
3. **Decide on Overall Correctness**
   - If the student's reasoning is **mostly consistent** with the correct reasoning and has **no major factual errors**, then it is considered **correct**.
   - If the student's reasoning **contains significant logical or factual errors**, or **omits critical steps** from the correct reasoning, then mark it **incorrect**.

Important Note: The student's reasoning does not have to match the official reasoning exactly; it just needs to be conceptually equivalent and free of serious contradictions.

Final Output Format:
- Provide your step-by-step analysis.
- At the end, write: **Final Answer**: "1" if you judge the student's reasoning is overall correct, "0" if it is overall incorrect.

Figure 10. Prompt for evaluation of model reasoning.

able to identify its direction and correctly answer the question. The ground truth explanation is the following:
"*The road markings indicate that the right lane can only be used for turning right. Therefore, I must position myself in the left lane to go straight or to the left.*"

Question Example 2

Human:
Unless explicitly stated otherwise, assume you are driving a car in France.
Select all correct answers to the following question from the available options. Detail your reasoning step by step based on road signs, markings, signals, and relevant driving rules. Then, provide the letters corresponding to your answer in the format: 'Answer(s): <letters>'
Question: This light is flashing, I must yield to this car:
Options: {A: Yes, B No.}

Question Output 1: QP-RBEA

QP-RBEA:
The relevant entities for this problem are: car [0.617, 0.664, 0.754, 0.798], traffic light [0.652, 0.398, 0.679, 0.521]. Reasoning: This car is stopped at the traffic light, so it does not have the right of way. Therefore, I do not need to yield to it while waiting for the light to turn green.
Answer(s): B.

Question Output 2: QPRB-EA

Human:
The relevant entities for this problem are: traffic light [0.682, 0.344, 0.717, 0.662], vehicle [0.623, 0.640, 0.805, 0.784], stop line marking [0.616, 0.721, 0.655, 0.804], stop sign [0.640, 0.585, 0.661, 0.628], traffic light [0.555, 0.603, 0.564, 0.635].
QPRB-EA:
Reasoning: The traffic light is malfunctioning, and the vehicle that wants to turn left must stop at the stop line marking. It is not a priority, so I do not have to yield to it. The stop sign is also present.
Answer(s): B.

In example 2, we compare providing the entities and bounding box coordinates as input (QPRB-EA), or letting the model detect them itself (QP-RBEA). Both models reach the same answer, but the second one's reasoning is wrong: it fails to identify a key entity (the stop line marking or the stop sign), leading to an incorrect reasoning even though the answer is correct. This highlights the importance of accurate entity detection in the reasoning process. The ground truth explanation is the following:
"*This traffic light is flashing, and there is no sign under the light. However, I can see that the vehicle on the right has a stop line marking and a stop sign. There is also another traffic light visible. Therefore, I do not have to yield to the right.*"

Question Example 3

Human:
Unless explicitly stated otherwise, assume you are driving a car in France.
Select all correct answers to the following question from the available options. Detail your reasoning step by step based on road signs, markings, signals, and relevant driving rules. Then, provide the letters corresponding to your answer in the format: 'Answer(s): <letters>'
Question: After these signs, will I be able to overtake this truck:
Options: {A: Yes, B No.}

Question Ouput 1: QP-EA

QP-EA:
Reasoning: The left lane is a storage lane that allows for left turns. It is strictly prohibited to use this type of lane for overtaking the truck. I therefore give up on making this maneuver.
Answer(s): B.

Question Ouput 2: QP-RB-RV-EA

QP-RB-RV-EA: The relevant entities for this problem are: speed limit sign [0.508, 0.294, 0.547, 0.401], truck [0.367, 0.304, 0.461, 0.546].
Human:
The relevant entities for this problem are: speed limit sign 

, truck .

QP-RB-RV-EA:
Reasoning: The prohibition signs must be read separately. The speed limit sign of 90 km/h applies to all road users. The second sign prohibits overtaking for vehicles transporting goods over 3.5 tons. Therefore, I will be able to overtake this truck.
Answer(s): A.

Example 3 shows a case where despite an imperfect entity detection ("speed limit sign" instead of "no overtaking sign", adding the image patches allows the model to correctly identify the important sign to answer the question. This is enabled by our cropping strategy that adds 70% of the image patch to increase the context and account for imperfect entity detection. On the contrary, the baseline QP-EA hallucinates the presence of a storage lane for turning left and answers incorrectly.

The ground truth explanation is: *This no overtaking sign prohibits goods transport vehicles from overtaking. It does not apply to me. Furthermore, I am driving on a one-way road with 2 lanes, so I can use the left lane to overtake. It is clear, as confirmed by a quick glance in the rearview mirror. I can therefore proceed with my maneuver while ensuring that I respect the speed limit of 90 km/h.*

D. Benchmarks comparison

Dataset	Real	Explanations	BBoxes	Relevant entities	Interleaved expl.	#Images	#QAs
Causal-VidQA [18]	✓	human-annotated	✓	✗	✗	27K	107K
ScienceQA [29]	∼	human-annotated	✗	✗	✗	10K	21K
PuzzleQA [5]	✗	LLM-generated	✗	✗	✗	2K	2K
GQA-CoT [41]	✓	LLM-generated	✓	single	✗	88K	88K
NuScenesQA [38]	✓	✗	✓	✗	✗	34K	460K
DriveLM-nuScenes [42]	✓	✗	✓	✗	✗	30K	443K
BDD-X [15]	✓	human-annotated	✗	✗	✗	7K	26K
LingoQA [30]	✓	LLM-generated	✗	✗	✗	28K	420K
DRIVINGVQA (Ours)	✓	human-annotated	✓	multiple	✓	4K	4K

Table 7. **Comparison of existing VQA datasets with explanations and/or for autonomous driving.** *Real* designed real-word images datasets. DRIVINGVQA uniquely combines real-world images from driving scenarios with (1) human-annotated, interleaved explanations and (2) multiple relevant entity annotations.