Towards Scenario Retrieval of Real Driving Data with Large Vision-Language Models

Tin Stribor Sohn[1*], Maximilian Dillitzer[1*], Lukas Ewecker[1], Tim Brühl[1], Robin Schwager[1], Lena Dalke[1], Philip Elspas[1], Frank Oechsle[1] and Eric Sax[2]
[1]*Dr. Ing. h.c. F. Porsche AG, Weissach, Germany*
[2]*Karlsruhe Institut für Technologie (KIT), Karlsruhe, Germany*
tin_stribor.sohn, maximilian.dillitzer1, lukas.ewecker, tim.bruehl, robin.schwager1, lena.dalke, philip.elspas,

Keywords: Large Vision-Language Models, Scenario Retrieval, Real Driving Data.

Abstract: With the adoption of autonomous driving systems and scenario-based testing, there is a growing need for efficient methods to understand and retrieve driving scenarios from vast amounts of real-world driving data. As manual scenario selection is labor-intensive and limited in scalability, this study explores the use of three Large Vision-Language Models, CLIP, BLIP-2, and BakLLaVA, for scenario retrieval. The ability of the models to retrieve relevant scenarios based on natural language queries is evaluated using a diverse benchmark dataset of real-world driving scenarios and a precision metric. Factors such as scene complexity, weather conditions, and different traffic situations are incorporated into the method through the 6-Layer Model to measure the effectiveness of the models across different driving contexts. This study contributes to the understanding of the capabilities and limitations of Large Vision-Language Models in the context of driving scenario retrieval and provides implications for future research directions.

1 INTRODUCTION

The automotive industry is undergoing a transformation driven by technological advances, particularly in the area of autonomous driving systems. As the complexity of vehicle functions rises, the need for manifold sensors and robust validation and testing methods becomes paramount. Traditional miles-driven approaches struggle to keep pace with the rapid evolution of autonomous driving technology and the complexity of real-world scenarios with rising automation levels. *Scenario-based testing* (SBT) has emerged as a promising solution to address the challenges associated with the validation of autonomous driving systems. By defining a comprehensive set of scenarios that encompass different driving conditions, environments, and edge cases, SBT provides a systematic approach to evaluate the performance and safety of autonomous vehicles. However, manually generating and selecting relevant scenarios can be time-consuming, resource-intensive, and limited in scalability. In recent years, the emergence of *Large Vision-Language Models* (LVLMs) has revolutionised the field of *artificial intelligence* (AI), enabling machines to understand and generate content across different modalities, including text and images. LVLMs, have the ability to understand complex scenes, objects, and contexts from both textual descriptions and visual input. Harnessing the power of LVLMs for *scenario retrieval* (SR) in the automotive industry has the potential to accelerate the validation process and increase test efficiency. By using LVLMs, automotive engineers and researchers can significantly reduce the time and effort required for scenario selection and validation. Focusing on the six layers of the *6-Layer Model* (6LM), three popular publicly available pretrained LVLMs, *Contrastive Language-Image Pretraining* (CLIP), *Bootstrapping Language-Image Pretraining 2* (BLIP-2), and BakLLaVA, are analysed. Quantitative and qualitative evaluations show the effectiveness and practicality of LVLMs in facilitating efficient and comprehensive SBT through SR.

[*] Equal contribution.

Sohn, T., Dillitzer, M., Ewecker, L., Brühl, T., Schwager, R., Dalke, L., Elspas, P., Oechsle, F. and Sax, E.
Towards Scenario Retrieval of Real Driving Data with Large Vision-Language Models.
DOI: 10.5220/0012738500003702

In *Proceedings of the 10th International Conference on Vehicle Technology and Intelligent Transport Systems (VEHITS 2024)*, pages 496-505
ISBN: 978-989-758-703-0; ISSN: 2184-495X

2 RELATED WORK

In the field of information retrieval in the automotive domain, research has mainly focused on objects, anomalies and scenarios.

Langner et al. (2019) propose a method for derivation of logical scenarios through clustering of dynamic length segments of driving data represented as time series. This provides the ability to derive distributions from clusters of concrete scenarios. A mapping between a functional description of a driving scenario and real driving data has not been elaborated.

Montanari et al. (2020) cluster recurring patterns of scenarios based on timeseries data. Clusters of similar scenarios and corner cases can be identified. This does not include the ability to query for these scenarios based on their functional descriptions.

Elspas et al. (2020) introduce a pattern matching mechanism based on regular expressions in order to extract driving scenarios from timeseries data. Rules for each scenario need to be derived in a knowledge driven process and adequate patterns need to be defined before the data is processed. As this rule-based method detects cut-ins and lane change maneuvers, it is possible to interpret extracted patterns with functional descriptions and encode them in retrievable representations.

In another contribution Elspas et al. (2021) have used fully *Convolutional Neural Networks* (CNNs) for time series in order to extract scenarios. As their approach requires labeled datasets with ground truth annotations for supervised learning, the applicability may be questioned and the domain may be limited regarding the complexity of annotating all relevant aspects in a representative training dataset.

Ries et al. (2021) propose a trajectory-based clustering method based on *Dynamic Time Warping* (DTW) for the identification of similar driving scenarios. This provides the ability to query similar trajectories of dynamic objects, but does not take into account all aspects of driving scenarios and abstraction layers.

To date, most of the methods for information retrieval include object retrieval methods, such as the works of Girshick et al. (2014); Girshick (2015); Kang et al. (2017); Hu et al. (2016). In the context of object retrieval, the work of Rigoll et al. (2023) needs to be specifically addressed, as it proposes a method using CLIP for object retrieval from automotive image datasets, combining the object labels into prompts. While it addresses object retrieval in the automotive domain for the purpose of querying objects for machine learning datasets, it does not address driving SR. To address more safety-critical driving scenarios, object retrieval can be extended by anomaly detection methods such as those provided by Unar et al. (2023); Rai et al. (2023).

The first to focus on the retrieval of overall driving scenarios is Wei et al. (2024). In this work, the authors propose a multi-modal *birds-eye-view* (BEV) retrieval method using BEV-CLIP, which provides a global feature perspective for holistic driving SR based on the overall context and layout. However, the method does not apply to a general 6LM-oriented framework, but to the occurrence and location of objects from BEV in complex scenes.

So far, no work has specifically addressed SR through functional scenario descriptions based on the structure of the 6LM, hence the ability to query and ground all aspects of driving scenarios with natural language.

3 THEORETICAL FRAMEWORK

3.1 Real Driving Data

Modern vehicles are equipped with manifold types of sensors in order to accurately perceive the environment. Additionally, software services provide value to the driver. Vehicle data can thus be recorded from different sources and in different modalities.

- Raw sensor data such as RADAR and LiDAR provide distance and velocity information through point clouds. Furthermore, camera sensors provide images and depth information from different perspectives of the ego-vehicle.
- Bus data includes all data transmitted over the vehicle's bus systems. It contains data from the entire functional chain, such as raw sensor data, as well as fused objects and high-level information.
- System log data is recorded directly from the different subsystems and digital in-vehicle services.

In scenario databases, retrieved data can be enriched with multiple external data sources such as map data or knowledge (Petersen et al., 2022). As LVLMs are mostly provided as pre-trained models for images and texts, the concept described in this work leverages camera data to extract the features for SR.

3.2 Scenario-Based Testing

The rising complexity of automotive systems makes it necessary to break down the real world into a subset of representative scenarios. SBT reduces the amount of *validation and verification* (V&V) while maintaining sufficient test coverage to achieve regulatory compliance. Scenarios are seen as the *"[...] temporal development between several scenes in a sequence of scenes"*, where a scene is a snapshot of the environment, including scenery, dynamic elements, and all self-representations of actors and observers as well as their relationships to each other (Ulbrich et al., 2015).

3.2.1 6-Layer Model

The work of Scholtes et al. (2021) provides context to these scenarios in the 6LM. The framework provides a structured description of driving scenarios, dividing them into six layers. The first layer describes the road network and its regulations, including road markings and traffic signs. To further detail its attributes for analysis purposes, the attribute-layers 1_1, 1_2, and 1_3 are created. Layer 1_1 contains only the road itself, while 1_2 contains the road markings and 1_3 the existing road signs. Layer 2 includes roadside structures, while layer 3 covers temporary modifications to layers 1 and 2, such as construction signs. Dynamic objects such as vehicles and pedestrians are introduced in layer 4 with a time-dependent description. Layer 5 is divided into attribute layers 5_1 and 5_2, which contain daytime and weather. Environmental conditions and digital information for communication are included in layer 6. In addition, to evaluate distinct queries, all six layers are further detailed in three levels, which can be seen in Table 1.

3.3 Scenario Descriptions

Menzel et al. (2018) introduce a terminology that outlines abstraction layers for driving scenarios, showing that scenarios take on different levels of abstraction at different stages of automotive system development along the V-model (Dröschel and Wiemers, 1999), as seen in Figure 1.

3.3.1 Functional Scenario

Functional scenarios are described in natural language during the concept and design phase of the development process, to be definable and understandable by human experts. They can contain different levels of detail and structures.

3.3.2 Abstract Scenario

Abstract scenarios provide a machine-interpretable format for execution using virtual validation techniques such as *X-in-the-Loop* (XiL). They are described using modeling languages or *Scenario Description Languages* (SDLs) (Bock and Lorenz, 2022).

3.3.3 Logical Scenario

Logical scenarios are described by parameter ranges and distributions rather than physical events. Virtual testing methods aim to sample from these distributions to generate concrete scenarios and evaluate them in test cases.

3.3.4 Concrete Scenario

Real driving data, consisting of sensor-, bus-, and system log data, represents concrete scenarios as it provides concrete physical values at specific points in time.

3.3.5 Test Case

Scenarios being mapped to metrics and acceptance criteria are called test cases. They can be functional, logical, or concrete scenarios. Hereby, validation metrics can be related to safety, comfort, and usability.

3.3.6 Relationship of Different Scenario Abstractions

In order to obtain logical scenarios, it is necessary to cluster concrete scenarios based on their specific attributes. The mapping of concrete and logical scenarios to functional scenarios makes them interpretable for human experts. This also applies to abstract scenarios in terms of machine interpretability. Test cases make them measurable. Drawing relationships between different abstraction layers is therefore a requirement for an effective SR method. (Figure 1).

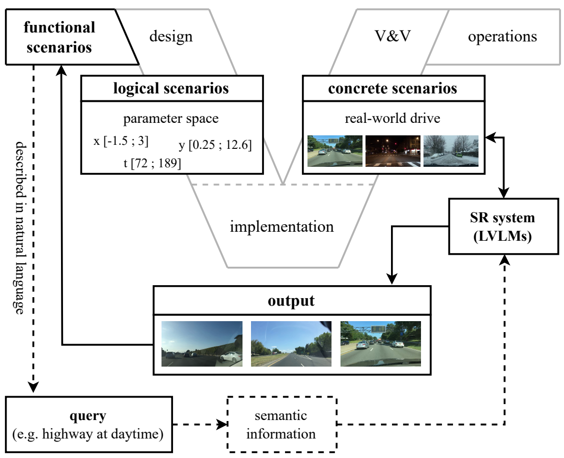


Figure 1: Concept of SR to map concrete scenarios to functional scenarios.

Table 1: Attribute Layers of the 6LM with: Road (1_1), Road Markings (1_2), Road Signs (1_3), Roadside Structures (2), Temporal Modifications (3), Objects (4), Daytime (5_1), Weather (5_2) and Communication (6), with up to three levels of detail.

		Level of Detail		
		1	2	3
Layers (6LM)	6	Source (e.g. Traffic Light)	Information (e.g. Color Red)	-
	5_2	Weather Condition (e.g. Rain)	Intensity (e.g. Strong)	-
	5_1	Illumination (e.g. Night)	Intensity (e.g. Dusk)	-
	4	Type (e.g. Pedestrian)	Behaviour (e.g. Moving)	Maneuvers (e.g. Cut-in)
	3	Type (e.g Construction)	Location (e.g. On-Road)	-
	2	Environment (e.g. Urban)	Scenery (e.g. Bridge)	Specification (e.g. Residential)
	1_3	Type (e.g. Street Sign)	Specification (e.g. Velocity)	Sign Value (e.g. 100km/h)
	1_2	Type (e.g. Lane)	Specification (e.g. Dashed)	Lane Count (e.g. 3)
	1_1	Category (e.g. Highway)	Road Character (e.g. Curvy)	Road Size (e.g. Large)

3.4 Scenario Retrieval

The urgency of safety assessment based on realistic driving scenarios, coupled with the open-world problem of automated driving, requires the collection of large amounts of driving data due to the multitude of use cases that need to be covered within *operational design domains* (ODDs) for higher levels of automation. Pütz et al. (2017) outlined a concept and motivation for a scenario database containing real-world driving scenarios for V&V. However, while this data is highly representative, it often lacks structure and requires additional annotation to be effectively queried for scenarios. These queries may involve defining parameter ranges, identifying patterns and trajectories, or applying similarity metrics. The challenge is to draw relationships between different scenario abstraction layers such as functional scenarios and real-world driving data. Traditional SR methods are inadequate to capture the complexity and variability of real-world scenarios. Using LVLMs for SR addresses these challenges by structuring and interpreting driving data to provide interpretable results for engineers.

3.5 Large Vision-Language Models for Information Retrieval

The ability to embed data in foundation models that have been pre-trained on large amounts of data with different modalities, such as speech, images, time series, or graphs, has gained significant interest in research, industry, and society. For application to specific tasks and data, there are two predominant approaches: fine-tuning and in-context learning. Fine-tuning involves updating the model weights to a specific target dataset and metric, which requires retraining the model. Due to the size of the parameters of such models, this process can require significant computational resources and time. In-context learning, on the other hand, does not require updating the weights. Instead, the goal is to provide the model with context for the specific task or dataset through targeted prompts. With these prompts, the model can generate more domain-specific responses. This can be done by manually exploring and designing prompts for the specific tasks, or by training additional models to generate prompts that achieve the best performance on the targeted task, called soft prompting (Lester et al., 2021). This method has shown superior performance to fine-tuning in resource-constrained downstream tasks (Devlin et al., 2019). To retrieve information, LVLMs perform an encoding process of both images and texts into a numerical representation that captures their semantic and visual features. After encoding both textual and visual inputs into feature representations, LVLMs employ algorithms to match and retrieve relevant information. These algorithms analyse the similarity between the encoded features of the query and the features of the database items. The retrieved information is then ranked based on the similarity scores, with the most relevant items presented to the user. Regardless of their potential, LVLMs are exposed to several challenges such as bias in data and fairness as well as the difficulty to distinguish right answers by the model and wrong but well formulated answers, also denoted as hallucination (Zhou et al., 2023).

4 METHOD

The method presented in this comparative study aims to provide a systematic approach to SR facilitated by functional scenario descriptions that effectively serve as natural language queries (Figure 1). Central to this method is the process of projecting images into the embedding space of a pre-existing LVLM. This operation involves encoding the images into vector representations within the model's embedding space. These vector representations are then stored and indexed in a vector database to allow efficient retrieval based on similarity metrics. To generate embeddings the LVLM is prompted with natural language queries. By leveraging the contextual understanding intrinsic in LVLMs, these embeddings encapsulate semantic information relevant to the queried scenarios. In addition, the retrieval process includes scoring the similarity between the query embeddings and those stored in a vector database. This similarity scoring mechanism facilitates the retrieval of the most relevant scenarios based on their proximity to the query in the embedding space. The retrieved scenarios are then returned as the output of the retrieval process.

4.1 Dataset

Some datasets, such as *Berkeley Deep Drive Explanation* (BDD-X) (Kim et al., 2018), contain a mapping of driving scenes to language descriptions. However, they do not encode structured information in the sense of the 6LM. Since no ground truth data is provided, a retrieval-precision-based evaluation approach is performed, which evaluates the relevance of the retrieved image with respect to the query. Therefore, the selection criterion of the dataset is that the driving scenes visually encode as much scenario-related information as possible. For evaluation purposes, the *Berkeley Deep Drive 100K* (BDD100K) dataset (Yu et al., 2020) is used. It contains 100,000 images of 1000 driving scenes in different contexts, seasons, daytime and weather conditions, taken from the ego-perspective of vehicle windshields. The variety of images related to all aspects of the 6LM including scenes on highways, rural roads, residential and urban areas, as well as various environmental conditions such as day, night, dusk, or dawn, makes it suitable for the evaluation of the presented method.

5 MODELS

To analyse the method proposed in this paper, the SR capabilities of three LVLMs, CLIP, BLIP-2, and BakLLaVA, are evaluated comparatively.

5.1 CLIP

CLIP is a multi-modal LVLM model capable of understanding images in the context of natural language (Radford et al., 2021). To be retrievable, all images are processed by the image encoder and projected into an embedding vector which is stored and indexed

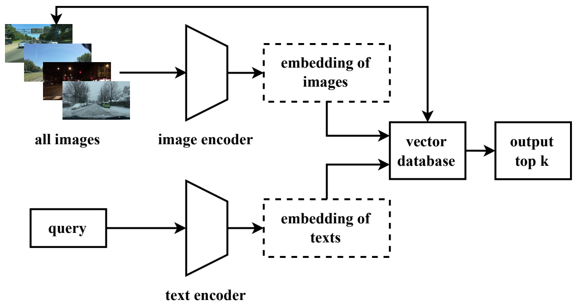


Figure 2: SR with CLIP.

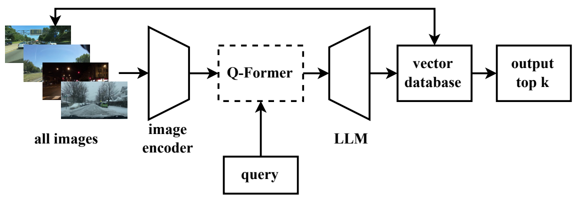


Figure 3: SR with BLIP-2.

within a vector database (Figure 2). Queries are then projected by the text encoder into the same embedding space in order to retrieve ranked results based on the cosine similarity for cross-modal understanding. As a result of using natural language queries, CLIP can search a large dataset of images and identify those that are relevant to the query, making it potentially suitable for SR.

5.2 BLIP-2

BLIP-2 is a generic and compute-efficient method for vision-language pre-training that leverages frozen pre-trained image encoders and *Large Language Models* (LLMs) (Li et al., 2023). Through the *Querying Transformer* (Q-Former), BLIP-2 is able to harvest the capabilities of already trained powerful vision and language models without having to update their weights when applied to downstream tasks such as visual question answering and image-text generation. Q-Former bridges the gap between two modalities and aligns their representation with improved performance, therefore showing potential for multimodal tasks such as SR (Figure 3).

5.3 BakLLaVA

BakLLaVA is based on the *Large Language and Vision Assistant* (LLaVA) model (Liu et al., 2023). LLaVA itself combines the *Large Language Model Meta AI* (LLaMA) model of Touvron et al. (2023) and a visual model using visual attributes. This architecture is extended by BakLLaVA using the Mis-

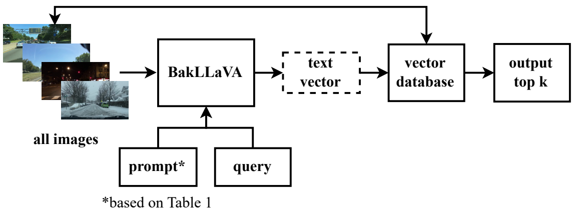


Figure 4: SR with BakLLaVA.

tral 7B LLM (Mistral AI, 2023), which improves the fusion of language and vision and refines the model's ability to understand and generate both text and images. It further enhances the capabilities of LLaVA by incorporating techniques for representation learning, attention mechanisms, and multi-modal fusion, resulting in a potentially suitable model for a more detailed SR. All images are processed by BakLLaVA with a prompt, instructing the model to caption the images based on the information described in Table 1. The captions are stored in a vector database with the index of the respective images to be queried for SR (Figure 4).

6 EVALUATION

To analyse the ability of SR, the LVLMs are evaluated considering each layer of the 6LM with its different levels of detail, as shown in Table 1. Consequently, the models are prompted using different queries corresponding to all attribute layers and their corresponding levels. A total of 189 queries is assigned to each model. Specifically, 60 queries are executed in layer 1, 24 in layer 2, 12 in layer 3, 36 in layer 4, 48 in layer 5, and 9 in layer 6. The imbalance in query distributions is due to the different attributes and granularities of each 6LM layer. The goal is to determine the effectiveness of the models in handling these queries, thereby revealing their capabilities of SR. Since labeled datasets of real driving scenarios are missing for SR, a manual analysis of the retrieved samples has to be performed. The *precision at k* (prec@k) metric provides a simple and clear interpretation of the results by focusing on the top recommendations of the model. For evaluation, the number of relevant items among the top k instances, denoted as n_k, determines the precision at a given value of k ($k \geq 1$). Since prec@k ranges from 0 to 1, it allows the model to reach up to 100% precision, especially for low values of k.

$$prec@k = \frac{n_k}{k} \tag{1}$$

For evaluation purposes, k is chosen to be $k = \{1; 5; 10\}$. Furthermore, the average of all precision values over all calculated prec@ks is used for simplification. This comprehensive evaluation framework aims to provide insight into the performance of the LVLMs, elucidate their capabilities across different layers and levels of detail, and provide a holistic understanding of their SR ability.

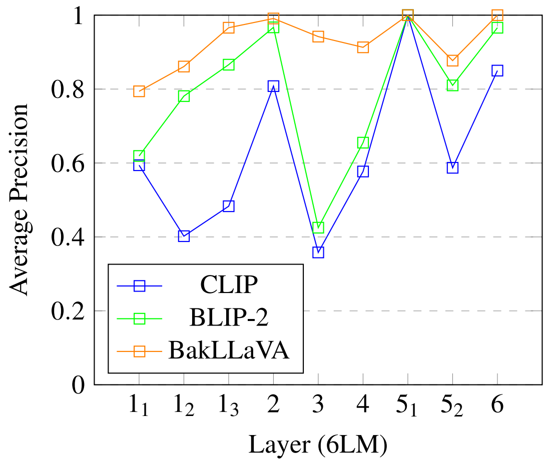
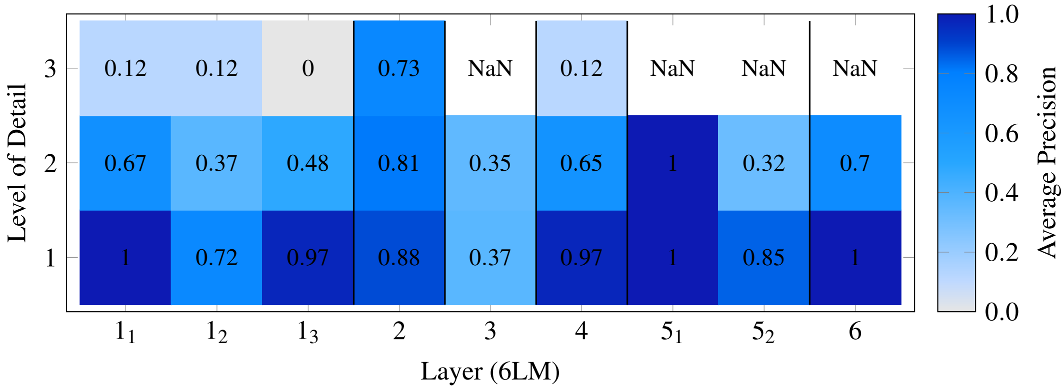


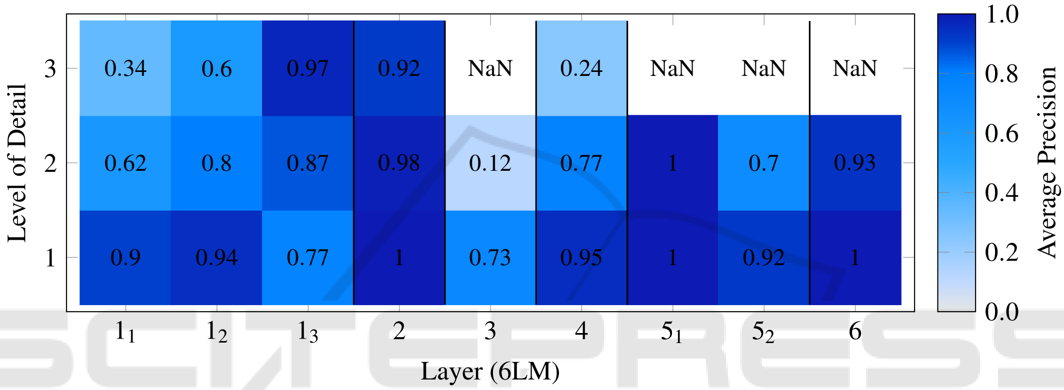
Figure 5: Average precision for CLIP, BLIP-2, and BakLLaVA on all layers, including attribute layers, of the 6LM.

6.1 Results and Discussion

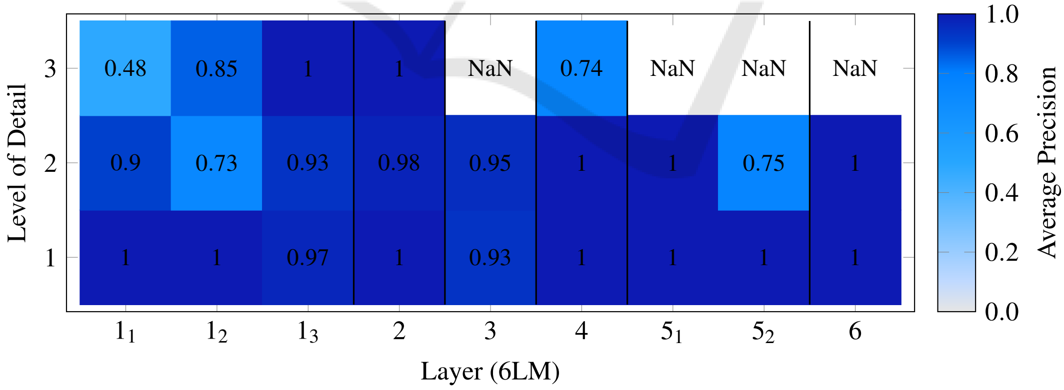
The evaluation of CLIP, BLIP-2, and BakLLaVA, shown in Figure 5, reveals differences in average precision across different layers of a scenario. BakLLaVA shows consistently higher average precision across all layers, achieving 92.74% compared to 78.78% achieved by BLIP-2 and 62.86% achieved by CLIP. The analysis identifies layer 5_1 as the best performing layer for all models with an average precision of 100%. Conversely, layer 3 has the lowest precision for CLIP and BLIP-2 with 35.83% and 42.50%, respectively, while layer 1_1 has the lowest precision for BakLLaVA with 79.44%. A spread in precision can be observed for all three models. CLIP and BLIP-2 have wider ranges of 64.17% and 57.50%, while BakLLaVA has a narrower range of 20.56%. All models show a slight trend indicating that precision tends to increase with higher layers. However, they differ in the increase or decrease trend of precision between certain layer transitions, especially from layer 1_1 to 1_2 and from layer 3 to 4. For all other layer transitions, the increase or decrease trend in precision is the same for the three models. Further examination of the attribute layers for layers 1 and 5 provides additional insight. For CLIP, there is no noticeable trend across the attribute layers of layer 1. In contrast, BLIP-2 and BakLLaVA show better identification of scenarios for the road signs in layer 1_3 compared to road markings



(a) CLIP



(b) BLIP-2



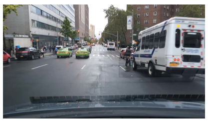
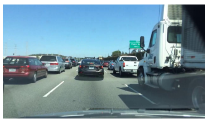
(c) BakLLaVA

Figure 6: Heatmaps showing average precision of the LVLMs with different levels of detail of the 6LM as in Table 1. NaN values indicate intentional omissions of experiments.

(1_2) and the road itself (1_1). Additionally, in layer 5, more detailed scenario queries for heavy and light weather conditions in attribute layer 5_2 lead to a decrease in precision.

To further evaluate the levels of detail, the heatmaps in Figure 6 show that CLIP achieves an average precision of 86.14% for level 1, 59.39% for level 2, and 21.67% for level 3 across all layers. Further, BLIP-2 achieves 91.27% for level 1, 75.43% for level 2, and 61.33% for level 3. In contrast, BakLLaVA achieves higher average precision across all levels, with 98.89% for level 1, 91.62% for level 2,

Table 2: Exemplary challenging queries and the corresponding answers of the BakLLaVA model.

Scenario	Query	Answer (BakLLaVA)
	What types of vehicles are visible in the scene?	In the scene, there are several types of vehicles present: A large white bus with blue stripes, which appears to be a public transit bus. Two green taxi cabs. A red car. A black SUV. A white car. A motorcycle. A truck. The vehicles are on a city street, and there are also pedestrians visible on the sidewalks.
	Where is the dark red car located?	The dark red car is in the left lane of the highway, moving in the same direction as the other traffic.
	In which directions are the different cars driving?	In the image, there are cars in the opposite direction of the camera's perspective. The most prominent vehicle is a yellow taxi cab, which is driving on the right side of the road. Behind the taxi, there are other vehicles, including a white van and a black SUV. The road appears to be a one-way street, and the vehicles are moving in the same direction as the taxi.
	Is a car turning to a specific direction in this scene?	The car in the image that appears to be turning to the right. It's a black car, and you can see the front end of it as it's making the turn. The car is positioned in the right lane, which is typically used for right turns, and it's angled towards the right side of the image, indicating that it's in the process of turning.

and 81.50% for level 3. Present *NaN* values indicate an intentional omission of an experiment for this particular level of detail, as no appropriate queries could be formulated. Dark coloring in the heatmap corresponds to high precision, with CLIP showing scattered coloring and achieving dark colors primarily at level 1. As the level of detail increases, the coloring becomes lighter, indicating lower precision. BLIP-2 still shows a scattered coloring with slightly darker patches compared to CLIP. Conversely, BakLLaVA's heatmap is dominated by dark colors at all levels, with only a few spots showing lighter coloring. All three models exhibit comparably low precision in layer 1_1 with detail level 3, especially when the model queried to distinguish between large and small intersections. In addition, layer 1_3 with detail level 3 shows a significant disparity in precision, with CLIP failing to detect speed signs with specified speed values, while BLIP-2 and BakLLaVA perform effectively in this regard. Layer 4 with detail level 3 poses a challenge for the models as they are queried to retrieve maneuvers such as following, braking, turning, and cut-ins. However, BakLLaVA shows a 62% and 50% higher precision in this task compared to CLIP and BLIP-2.

To further investigate BakLLaVA's ability to address specific aspects and levels of detail, as it is the best performing LVLM in this evaluation, a qualitative analysis was performed using an image dialogue. Table 2 shows four exemplary dialogues out of a set of 25 questions that were asked to further investigate the possible level of detail of additional aspects. Red colored answers represent wrong answers of the model to the given query. The results show that BakLLaVA is able to answer the query adequately even for fine-grained scene descriptions such as color, number and location of certain objects. Boundaries were especially investigated in the description of motion directions, including car directions, turning directions, and lane directions. In these cases, model hallucination was observed, in which the model not only provided incorrect movement directions, but also began to describe the scene incorrectly.

The evaluation results show the general ability of LVLMs to query camera data of real driving data for driving scenario related information. Up until the 10 best recommendations by the model ($k = 10$), CLIP, BLIP-2 and BakLLaVA are able to achieve high average precision scores. With higher degrees of detail in the query, the performance of both, CLIP and BLIP-2, degrades, while BakLLaVA is still able to encode many contexts in further detail degrees, such as describing the exact speed limit value on a speed sign.

Further investigations through a visual dialogue show, that BakLLaVA can provide even more detailed information about object attributes and locations. On the other hand, all of the evaluated models struggle with information encoded in temporal sequences of images such as detailed maneuvers or object movement directions embedded in layer 4. This might also be related to the static nature of the inputs, as not videos but single input images were fed into the models.

7 CONCLUSION

The elaborated method outlines the potential of using pre-trained LVLMs for semantic enrichment and retrieval of real driving data with natural language queries in the form of functional scenario descriptions. Specifically, BakLLaVA, consisting of an image encoder and Mistral 7B as the LLM backbone, achieves accurate query results even for detailed specifications such as the location and color of objects encoded in the images.

Future work should focus on several key areas. One key is to create a dataset tailored for SR with LVLMs that includes multi-modal driving data such as time series or point clouds additional to images. Incorporating external data sources such as map- and weather data can provide additional semantic structure to produce meaningful joint embeddings. The ability of LVLMs to incorporate other SR tasks, such as querying abstract scenario descriptions from concrete, logical, and functional scenarios, offers potential for more efficient and effective SBT. Metrics like *recall at k* (recall@k) should be evaluated in addition to prec@k to ensure the relevance of the retrieved scenarios. Furthermore, future research should investigate prompt engineering techniques, incorporate taxonomies for different use cases, and explore the temporal domain using video language models. The impact of fine-tuning compared to in-context learning, and the associated trade-off in computational cost for the SR task, may have important implications for future research directions. User studies with domain experts querying scenarios can be conducted to explore the feasibility of the concept and the ability of the models to cope with domain-specific language. Analysing combined queries that jointly integrate different scenario layers can provide a more comprehensive understanding of the SR capability. Besides retrieval performance, additional metrics such as computational efficiency, storage requirements, and retrieval time should be considered. These efforts will advance SR methods in the automotive domain for V&V tasks.

REFERENCES

Bock, F. and Lorenz, J. (2022). Abstract natural scenario language version 1.0.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dröschel, W. and Wiemers, M. (1999). *Das V-Modell 97: der Standard für die Entwicklung von IT-Systemen mit Anleitung für den Praxiseinsatz*. Oldenbourg Wissenschaftsverlag, Berlin, Boston, 2014th edition.

Elspas, P., Klose, Y., Isele, S. T., Bach, J., and Sax, E. (2021). Time series segmentation for driving scenario detection with fully convolutional networks. In *VEHITS*, pages 56–64.

Elspas, P., Langner, J., Aydinbas, M., Bach, J., and Sax, E. (2020). Leveraging regular expressions for flexible scenario detection in recorded driving data. In *2020 IEEE International Symposium on Systems Engineering (ISSE)*, pages 1–8. IEEE.

Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587.

Hu, R., Xu, H., Rohrbach, M., Feng, J., Saenko, K., and Darrell, T. (2016). Natural language object retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4555–4564.

Kang, K., Li, H., Xiao, T., Ouyang, W., Yan, J., Liu, X., and Wang, X. (2017). Object detection in videos with tubelet proposal networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 727–735.

Kim, J., Rohrbach, A., Darrell, T., Canny, J., and Akata, Z. (2018). Textual explanations for self-driving vehicles. In *Proceedings of the European conference on computer vision (ECCV)*, pages 563–578.

Langner, J., Grolig, H., Otten, S., Holzäpfel, M., and Sax, E. (2019). Logical scenario derivation by clustering dynamic-length-segments extracted from real-world-driving-data. In *VEHITS*, pages 458–467.

Lester, B., Al-Rfou, R., and Constant, N. (2021). The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Li, J., Li, D., Savarese, S., and Hoi, S. (2023). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. (2023). Visual instruction tuning.

Menzel, T., Bagschik, G., and Maurer, M. (2018). Scenarios for development, test and validation of automated vehicles. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1821–1827.

Mistral AI (2023). Mistral 7b: The best 7b model to date, apache 2.0. Mistral AI News. Accessed: February 22, 2024.

Montanari, F., German, R., and Djanatliev, A. (2020). Pattern recognition for driving scenario detection in real driving data. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 590–597. IEEE.

Petersen, P., Stage, H., Langner, J., Ries, L., Rigoll, P., Hohl, C. P., and Sax, E. (2022). Towards a data engineering process in data-driven systems engineering. In *2022 IEEE International Symposium on Systems Engineering (ISSE)*, pages 1–8. IEEE.

Pütz, A., Zlocki, A., Bock, J., and Eckstein, L. (2017). System validation of highly automated vehicles with a database of relevant traffic scenarios. *situations*, 1:E5.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision.

Rai, S. N., Cermelli, F., Fontanel, D., Masone, C., and Caputo, B. (2023). Unmasking anomalies in road-scene segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4037–4046.

Ries, L., Rigoll, P., Braun, T., Schulik, T., Daube, J., and Sax, E. (2021). Trajectory-based clustering of real-world urban driving sequences with multiple traffic objects. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 1251–1258.

Rigoll, P., Langner, J., and Sax, E. (2023). Unveiling objects with sola: An annotation-free image search on the object level for automotive data sets. *arXiv preprint arXiv:2312.01860*.

Scholtes, M., Westhofen, L., Turner, L. R., Lotto, K., Schuldes, M., Weber, H., Wagener, N., Neurohr, C., Bollmann, M. H., Körtke, F., et al. (2021). 6-layer model for a structured description and categorization of urban traffic and environment. *IEEE Access*, 9:59131–59147.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023). Llama: Open and efficient foundation language models.

Ulbrich, S., Menzel, T., Reschka, A., Schuldt, F., and Maurer, M. (2015). Defining and substantiating the terms scene, situation, and scenario for automated driving. In *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, pages 982–988.

Unar, S., Su, Y., Zhao, X., Liu, P., Wang, Y., and Fu, X. (2023). Towards applying image retrieval approach for finding semantic locations in autonomous vehicles. *Multimedia Tools and Applications*, pages 1–22.

Wei, D., Gao, T., Jia, Z., Cai, C., Hou, C., Jia, P., Liu, F., Zhan, K., Fan, J., Zhao, Y., et al. (2024). Bev-clip: Multi-modal bev retrieval methodology for complex scene in autonomous driving. *arXiv preprint arXiv:2401.01065*.

Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., and Darrell, T. (2020). Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645.

Zhou, Y., Cui, C., Yoon, J., Zhang, L., Deng, Z., Finn, C., Bansal, M., and Yao, H. (2023). Analyzing and mitigating object hallucination in large vision-language models.