Embodied Understanding of Driving Scenarios

Yunsong Zhou[1,2], Linyan Huang[1], Qingwen Bu[1,2], Jia Zeng[1], Tianyu Li[1], Hang Qiu[3], Hongzi Zhu[2(✉)], Minyi Guo[2], Yu Qiao[1], and Hongyang Li[1(✉)]

[1] OpenDriveLab at Shanghai AI Lab, Shanghai, China
[2] Shanghai Jiao Tong University, Shanghai, China
[3] University of California, Riverside, USA

Abstract. Embodied scene understanding serves as the cornerstone for autonomous agents to perceive, interpret, and respond to open driving scenarios. Such understanding is typically founded upon Vision-Language Models (VLMs). Nevertheless, existing VLMs are restricted to the 2D domain, devoid of spatial awareness and long-horizon extrapolation proficiencies. We revisit the key aspects of autonomous driving and formulate appropriate rubrics. Hereby, we introduce the Embodied Language Model (ELM), a comprehensive framework tailored for agents' understanding of driving scenes with large spatial and temporal spans. ELM incorporates space-aware pre-training to endow the agent with robust spatial localization capabilities. Besides, the model employs time-aware token selection to accurately inquire about temporal cues. We instantiate ELM on the reformulated multi-faced benchmark, and it surpasses previous state-of-the-art approaches in all aspects. All code, data, and models are accessible at https://github.com/OpenDriveLab/ELM.

1 Introduction

Embodied understanding enables intelligent agents (*e.g.*, self-driving vehicles, robots, and drones) to interpret instructions and analyze scenes based on their experience [28,93]. However, this critical but challenging task is yet to be solved. Recently, benefiting from their extensive knowledge and causal reasoning capability, vision language models (VLMs) [3,48,53,98] have achieved remarkable progress in general vision [10,46,49,56,61]. The utilization of VLMs provides a question-answering framework to engage with a scene and contribute to common sense comprehension. When it comes to driving scenarios, embodied approaches via VLMs have the potential to surpass both rule-based [25,70,72,82] and data-driven learning-based [9,15,33,34] methods in unforeseen scenarios [12,53,93].

Y. Zhou, L. Huang and Q. Bu—Equal contribution.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-73033-7_8.

A. Leonardis et al. (Eds.): ECCV 2024, LNCS 15120, pp. 129–148, 2025.
https://doi.org/10.1007/978-3-031-73033-7_8

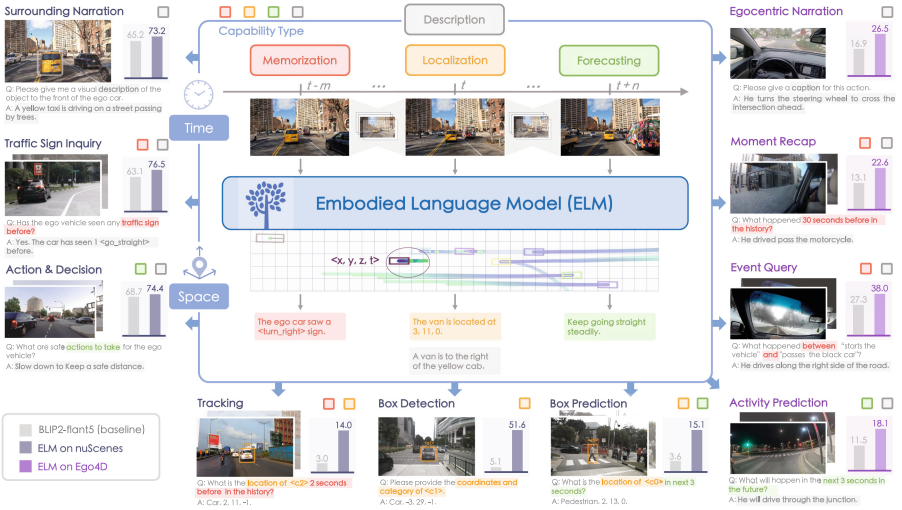

Fig. 1. ELM is an embodied language model for understanding the long-horizon driving scenarios in space and time. Compared to the vanilla vision-language model (VLM) being confined to the scene description task, we expand a wide spectrum of new tasks to fully leverage the capability of large language models in an embodiment setting. ELM achieves significant improvements in various applications.

To cope with complex driving scenarios, it is crucial for an embodied agent to obtain a complete 4D scene understanding, particularly in extensive spatial scale and extended temporal duration. As depicted in Fig. 1, this calls for four pivotal capabilities, including **1)** *description*: the agent is able to describe the surrounding environments; **2)** *localization*: rather than merely assessing approximate position, the agent needs to pinpoint a particular object in the 3D space; **3)** *memorization*: the agent needs to retrieve specific events that have occurred; and **4)** *forecasting*: the agent is required to foresee a certain future from the given history.

Recently, attempts are conducted to incorporate VLMs into the autonomous driving domain. Current methods are instrumental in crafting narrations encompassing the surroundings environment [58], traffic participants [16], road components [18], potential interactions [42,89], and driving behaviors [43,71,91]. Nevertheless, the capabilities of vanilla VLMs are limited to generating narrative phrases, namely description. Their sense of space and time remain unexplored, as existing works can only describe rough position information [68] and achieve information retrieval in a short period [46,49]. As such, the absence of localization, memorization, and forecasting refrains VLMs from the embodied understanding of driving scenarios.

To this end, we introduce **Embodied Language Model (ELM)** for the proposed driving scene understanding problem. As highlighted in Fig. 1, in contrast to conventional VLMs which only possess description capability, ELM

extends the capabilities of language models in large spatial and temporal horizons. Amongst the newly formulated problem, a suite of tasks and evaluation protocols are presented. The key challenges are presented as follows:

Long Horizon in Space. Since VLM decoders are natrually insensitive to numbers, an intuitive solution would be to rewrite the vocabulary [4] and pre-train models on numerically relevant tasks with the replaced words. However, excessive training on a single type of data may lead to catastrophic forgetting [95]. Data diversity is therefore of crucial importance. We propose *space-aware pre-training* along with a diverse data collection and auto-labeling process. We orchestrate over 3,000 h of data and 9 million pairs of diverse annotations from the open world, incorporating the public autonomous driving datasets nuScenes [8] and Waymo [78], the internet-derived dataset YouTube and the egocentric dataset Ego4D [28]. This enables the autonomous agent to acquire spatial localization competence while preserving the initially robust descriptive aptitudes.

Long Horizon in Time. Summarizing long historical time-series data is computationally burdensome with significant redundancy. A straightforward way is to split and sample the video into images [46,49,61]. While there is an attempt to summarize a film as a sequence of chronologically occurring events [77], it does not allow the agent to recall events from a brief moment in a lengthy video. We are of the opinion that the crux lies in enabling the agent to efficiently retrieve the most pertinent content from long-term memory based on the given instruction. To accomplish this, we opt in a module named *time-aware token selection*. The module encodes each frame into sparse tokens and builds a token bank. A set of learnable queries is leveraged to extract the most relevant moment-specific and content-specific cues emphasized in the instruction, enabling effective long-term information retrieval.

Benchmark. To evaluate ELM and other VLMs, we assemble a new evaluation suite comprising ten distinct tasks. These tasks encompass evaluations of both individual and integrated competencies in description, localization, memorization, and forecasting, as delineated in Table 1. The devised tasks include descriptive tasks within the purview of vanilla VLMs, as well as tasks involving spatio-temporal localization and dynamic information prediction. While the primary focus of this investigation pertains to driving scenarios, it is worth noting that the incorporation of daily indoor scenarios can serve as a valuable means to assess VLMs' capacity for long-term event reasoning. The details of the formulated tasks are described in Sect. 2.

The **contributions** are three folds: **a)** We revive driving scene understanding by delving into the embodiment philosophy. This involves a deconstruction of its definition and basic capabilities, along with a series of novel tasks and a comprehensive evaluation benchmark. **b)** We propose ELM, a vision-language model for embodied understanding in driving scenarios. Our proposed space-aware pre-training strategy and time-aware token selection enhance agents' comprehension in long-range four-dimensional space. **c)** We validate ELM on the all-encompassing tasks for cross-domain scenarios. Experimental results demon-

strate the superiority of our method compared to LLaMA-Adapter V2 [27], LLaVA [53], Otter [46], VideoChat [49], *etc.*. Figure 1 visualizes the achieved improvement compared to BLIP2-flant5 [48] across ten distinct tasks.

Table 1. Performing Tasks for Embodied Understanding of Driving Scenarios. We supplement the evaluation of long-term memory with long videos from Ego4D [28], which is lacking in self-driving datasets. The gray-colored tasks are already applicable to vanilla VLMs. `S`: the span in space; `R`: the resolution in space; `T`: total duration; `F`: the number of frames; `#`: the number of QA pairs.

Tasks	Fine-tune Dataset	Capability				Statistics		
		Description	Localization	Memorization	Forecasting	S(m)/R(m)	T (s)/F	#
Surrounding Narration	nuScenes [8]	✓				30/5	0.5/1	142K
Traffic Sign Inquiry		✓		✓		30/1	3.5/7	20K
Action & Decision		✓			✓	30/5	3.5/7	301K
Box Detection			✓			50/1	0.5/1	232K
Tracking			✓	✓		50/1	3.5/7	131K
Box Prediction			✓		✓	50/1	3.5/7	133K
Egocentric Narration	Ego4D [28]	✓				20/3	3/1	357K
Moment Recap		✓		✓		20/3	60/20	70K
Event Query		✓		✓		20/3	60/20	70K
Activity Prediction		✓			✓	20/3	60/20	69K

2 Problem Setup

Based on the analysis of the pivotal competencies involved in embodied understanding, the newly proposed benchmark thoroughly evaluates VLMs from the perspective of description, localization, memorization, and forecasting. Utilizing the nuScenes [8] and Ego4D [28] datasets, we formulate ten question-answering (QA) tasks as listed in Table 1.

Built on top of the nuScenes dataset, we present three tasks which are for prompting embodied agents to provide descriptions of the current scene, recall previously observed traffic elements, and predict future states. Furthermore, we devise three localization-related tasks, which require embodied agents to deduce the 3D positions of 2D query points in the present, past, and future. Completing these positioning tasks necessitates robust spatial perception and temporal reasoning capabilities. To ensure that VLMs remain unbiased towards driving scenes, we incorporate the Ego4D dataset for evaluation in common scenarios. These tasks require the description of ongoing events, inquiry of past events, and prediction of future events. The scenes in Ego4D consist of prolonged videos, and this necessitates a greater understanding over long time spans.

The formulation of each task is elaborated as follows:

– *Surrounding Narration:* providing an overall description of the surroundings, namely attribute, presence, and movement of traffic objects on a single frame.

- *Traffic Sign Inquiry:* identifying and recalling traffic signs and lane markings observed within 3.5 s in the past.
- *Action & Decision:* providing a high-level planning-related instruction to foresee potential interactions and make driving decisions.
- *Box Detection:* inferring the 3D coordinate and category based on the 2D query point on a single frame.
- *Tracking:* retrieving the 3D trajectory and category of the object queried by the 2D pixel position of the current frame for the last 3.5 s.
- *Box Prediction:* inferring the future 3D location and category of a queried object given its current 2D coordinate and 3.5 s past observations.
- *Egocentric Narration:* describing self-behaviors (actions and interactions with the surroundings) based on an egocentric single-frame input.
- *Moment Recap:* indicating an event that occurred at a specific point in time within the last 60 s.
- *Event Query:* deducing the content of an specific event through the analytical examination of its antecedent and subsequent events in a 60-second video.
- *Activity Prediction:* predicting an event that will happen at a designated future moment in a 60-second video.

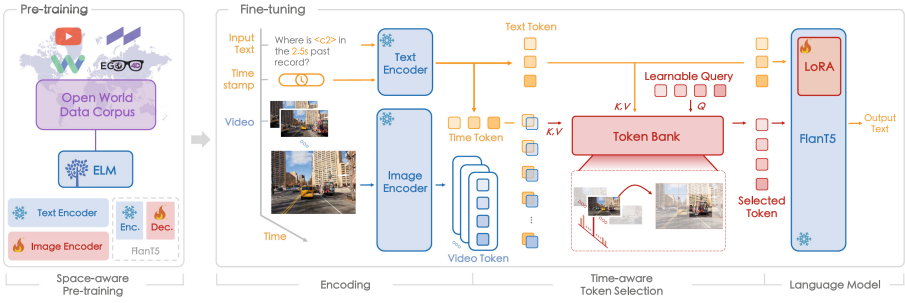

Fig. 2. Systematic Pipeline of ELM. It consists of Pre-training by open-world data corpus and Fine-tuning on diverse tasks. To initialize the Space-aware Pre-training, we collect extensive image-text pairs from the world, empowering ELM with spatial localization while preserving the description ability in driving scenarios. In the fine-tuning process, the inputs to ELM are videos, timestamps, and text prompts. After encoding the inputs into tokens, ELM leverages the proposed Time-aware Token Selection to gather the appropriate tokens as instructed by prompts. Finally, the tokens are sent to the language model to generate output texts.

In contrast to previous datasets and tasks [42,43,58,68,71,76,89], the proposed benchmark incorporates both spatial and temporal evaluation, necessitating embodied agents to have a correct understanding of the complex driving scenes. We set up this benchmark for assessing embodied understanding in driving scenarios and harmonizing diverse driving-related objectives.

License and Privacy Considerations. All the annotated data (benchmarks and open-world data corpus mentioned in Sect. 3.2) comply with the CC BY-NC-SA license. Following [2,39,90,97,99], we safeguard the rights of the data owners and prevent privacy leakage by distributing redirection links instead of publishing image contents. Personal identification information will not be leaked to the public. For more details about license and privacy, please consult the Appendix.

3 Methodology

3.1 Overview

We aim at enhancing agents' spatial perception with diverse pre-training data and dealing with long time series through adaptive token selection. Figure 2 illustrates the architecture of our framework. ELM begins with Space-aware Pre-training (Sect. 3.2) on image-text pairs. During this phase, ELM focuses on the vocabulary related to spatial understanding and learns a robust visual encoder through extensive training. Throughout the fine-tuning across varied tasks, all encoders of ELM are frozen. In the Encoding process, the text prompt and timestamp are encoded by BERT [17], while each video frame is transformed into fixed-length feature tokens using the EVA model [26]. In Time-aware Token Selection (Sect. 3.3), the video tokens are fed into a token bank along with the text and timestamp tokens, and the token bank adaptively selects the desired tokens based on the text prompt. Lastly, the FlanT5 [69] model, fine-tuned with LoRA [31], generates the output text to tackle various tasks in our benchmark.

Table 2. Statistics of pre-training data and comparison of data collection with other VLMs. Our pre-train data surpasses that in general vision (top) and autonomous driving (middle) in terms of quantity and diversity. `Anno`: the type of annotations; `Des`: description; `Loc`: localization.

Method	Pre-train Data	# Frames	Duration (hours)	Geographic Countries	Diversity Cities	Anno
LLaVA [53]	COCO [52]	150K	–	–	–	Des
VideoChat [49]	Self-Collected	18K	–	–	–	Des
Vid-ChatGPT [61]	ActivityNet-200 [7]	100K	–	–	–	Des
nuScences-QA [68]	nuScenes [8]	460K	5.5	2	2	Des
DriveGPT4 [92]	BDD-X [43]	28K	77	1	4	Des
LLM-driver [12]	Self-Collected	160K	-	-	-	Des
DriveLM [76]	nuScenes [8], CARLA [21]	188K	95	3	3	Des
ELM (Ours)	nuScenes [8]	7.4M	5.5	2	2	Des, Loc
	Waymo [78]	450K	6.4	1	6	Des
	YouTube	1.1M	1474	$\geq$40	$\geq$ 709	Des
	Ego4D [28]	300K	1638	9	74	Des

3.2 Space-Aware Pre-training

Open World Data Collection. In pursuit of spatial localization while retaining the description capacity for driving scenarios, we collect an open-world data corpus for the space-aware pre-training. As depicted in Table 2, the data corpus is derived from a variety of sectors. Representative datasets for autonomous driving, such as nuScenes [8] and Waymo [78], constitute our fundamental resources. These two datasets comprise a total of 11.9 h of data, capturing scenes from five different cities: Boston, Singapore, San Francisco, Phoenix, and Mountain View. YouTube, renowned for its extensive data and diverse content, serves as a critical resource for our research. We collect a total of 1,474 h of publicly available videos from over 709 cities in more than 40 countries using web crawlers. The collected data covers a wide range of locations, including urban areas, rural regions, and various weather conditions. For a broader vision, we utilize the Ego4D dataset [28], which provides an in-depth understanding of daily activities worldwide. There are 931 camera wearers contributing a total of 1,638 h of footage from 74 cities. We aggregate an extensive and diverse dataset for pre-training, which goes far beyond those adopted in other VLMs [12,49,53,61,68,92].

Auto-labeling with Human in the Loop. With the objective of enhancing models' spatial comprehension, we design a localization labeling process based on nuScenes [8] in Fig. 3. To ensure the diversity of questions, we use GPT-4 [63] to generate massive unique templates for text prompts. In response to the GPT's instability, we execute a manual selection process to assemble a set of 1000 high-quality templates. Regarding the location ground truth labels, we leverage the point clouds and camera parameters to establish the correspondence between 2D pixels and 3D point coordinates. In addition, we employ density-based point sampling to achieve uniform coverage in 3D space, followed by a rule-based method to assign the labels to the templates. Collectively, we create a total of 7.4M QA pairs about location. The pipeline is detailed in the Appendix.

For preventing catastrophic forgetting during pre-training [95], we introduced a large number of description labels into the data corpus, thereby enhancing the diversity of the dataset. The right side of Fig. 3 illustrates our description labeling pipeline, and the labels include descriptive sentences of the overall scenario, transport elements, and driving decisions. Particularly, two rounds of quality check are implemented to maintain a high standard of labeled data. The annotation pipeline starts by removing noisy, interfering, and blurry images from `Node` 1 to `Node` 3. After crawling raw data from the open world, the inspectors extract 10% of a batch of images for a quality check to determine if the batch should be retained. The image selection process primarily involves sorting out the worst N samples in terms of quality from a quantitative set of video clips based on standards like lighting, resolution, and clarity. These are then returned to the reserve pool, with the remainder forwarded to the next process. If a video source is repeatedly flagged as poor quality, it is placed on a blacklist. The qualified data batches are fed into LLaMA-Adapter V2 [27] to generate caption batches, while others are discarded. Following this, the second quality

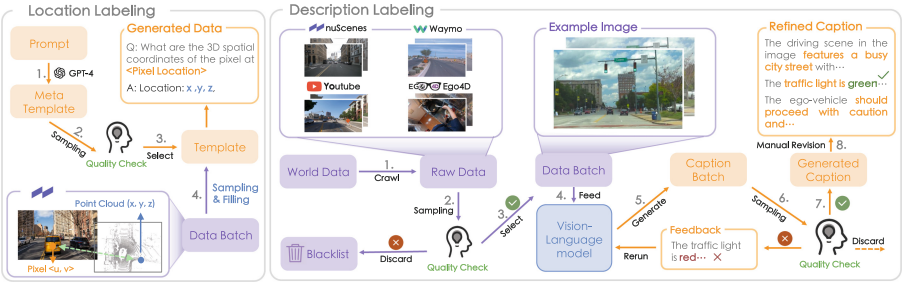

Fig. 3. Annotation workflow with human quality check in the loop. For location labeling: we first select diverse templates from the GPT generated candidates. Pixel-point pairs as annotated in the nuScenes [8] are then sampled and filled into the templates to form our location pre-training data. **For description labeling:** `Node` 4 utilizes LLaMA-Adapter V2 [98] to obtain diverse labels on nuScenes, Waymo [78], YouTube, and Ego4D [28] with predefined prompts. Two rounds of quality check are conducted in `Node` 3 and 7 by inspectors to guarantee the image and caption quality.

check on the generated caption is performed in `Node` 6–7. The revised captions will be saved as the final annotations in `Node` 8. In instances where a data batch fails to meet quality standards, inspectors will furnish feedback to the model for the purpose of refining the generated captions in subsequent iterations. Labeling details, discarded images, and annotation examples are in the Appendix.

Following the workflow above, we have amassed over 9 million annotations, indicating a substantial increase in the scale and diversity of the dataset used for pre-training. The comparison of annotation quality and diversity will be further demonstrated in Sect. 4.3 and the Appendix.

Tokenizer. It is argued that VLMs are insensitive to numbers [22]. An RT2-like tokenizer [4] is implemented to enable a general VLM to perform location prediction in the form of text. We divide the 3D space into 1-meter resolution grids and quantify the position of the target point as the index of the grid. Then we rewrite the least frequently used words in FlanT5 to represent the grid index, which is referred to as space-relevant vocabulary. Hence the 3D localization could be deemed as a language modeling task.

3.3 Time-Aware Token Selection

To effectively memorize and forecast events in long time-series videos, it is essential to encode the scene using a timestamp-sensitive representation and select tokens wisely. Thus, we introduce the Time-aware Token Selection module, which utilizes the input text prompt as guidance to select a fixed number of relevant tokens from the video. These selected tokens are then incorporated into the language model as visual input. To facilitate interaction among videos, timestamps, and prompts, it is important to align their embeddings within the textual feature space, for which we perform the following design.

Video Encoding. Initially, we utilize Q-former [48] to align the video features with language model inputs:

$$\begin{aligned} q_v^t &= \texttt{SA}(F_v^t) \in \mathbb{R}^{32\times d}, \\ \hat{q_v}^t &= \text{Q-Former}(q_v^t, q_l) \in \mathbb{R}^{32\times d'}, \end{aligned} \tag{1}$$

where q_v^t and $\hat{q_v}^t$ denote the video tokens before and after Q-former at timestamp t, $F_v^t \in \mathbb{R}^{HW\times d}$ represents the video frame feature at timestamp t generated from the visual encoder (*i.e.*, EVA [26]), $q_l \in \mathbb{R}^{32\times d}$ is a group of learnable embeddings used to transform video content into textual information [48], and we use d and d' to denote the dimension of visual and textural embeddings, respectively. $\texttt{SA}(\cdot)$ is the Slot Attention module [54] to acquire visual representations while further reducing redundancy.

Timestamp Encoding. Conventional techniques like sinusoidal [80] or learnable encoding [20] of timestamps mismatch with language models since they're from different domains. In contrast, we propose to transform timestamps into the form of text. Subsequently, we leverage the FlanT5 [69] encoder for generating embeddings aligned with temporal information contained in the input text queries. Our approach skillfully circumvents the challenge of aligning temporal encoding with the text embedding space.

Adaptive Selection via Token Bank. The key to token selection lies in enabling the model to comprehend timestamps and video content, thereby identifying the most relevant tokens to the given prompt within the time series. In pursuit of this, we introduce the token bank module, which leverages the weighted aggregation of tokens to dynamically preserve both query-specific and overall contextual information. Specifically, we initiate the process by creating a set of learnable queries, represented as $q_i \in \mathbb{R}^{n\times d'}$. Employing a cross-attention mechanism, these learnable queries effectively comprehend the input prompt, with the concatenation of timestamps and visual embeddings serving as keys within the cross-attention module. Meanwhile, the learnable queries play the role of extracting the most relevant visual features E_{vis} from a long-time series:

$$\begin{aligned} q_{\text{mid}} &= \texttt{MHCA}\Big[q_i, \text{T5}_{\text{Enc}}(T_p), \text{T5}_{\text{Enc}}(T_p)\Big], \\ E_{\text{vis}} &= \texttt{MHCA}\Big[q_{\text{mid}}, \texttt{concat}\left(\hat{q_v}, \text{T5}_{\text{Enc}}(T_t)\right), q_v\Big], \end{aligned} \tag{2}$$

where T_p and T_t represent the text prompt and timestamp, respectively. q_v and $\hat{q_v}$ correspond to the entire video representation before and after Q-former, while q_{mid} serves as an intermediate token that incorporates textual prompt. $\texttt{MHCA}(\cdot)$ denotes multi-head cross attention and $T5_{\text{Enc}}$ is the FlanT5 encoder [69].

The selected visual features E_{vis} will then be processed by Q-former and fed into the language model as the visual embedding. As queries and keys in the cross attention are aligned within the textual domain, our approach effectively identifies and extracts moment- and content-specific visual representations. A more detailed illustration of the pipeline is given in the supplementary materials.

4 Experiments

The fine-tuning datasets of all ten tasks are built upon nuScenes [8] and Ego4D [28]. Additional information (annotations, dataset statistics, implementation details, training strategies, *etc.*) is provided in the supplementary materials.

Evaluation Metrics. For localization-related tasks, *i.e.*, Tracking, Box Detection, and Box Prediction, we propose metrics specifically designed for the assessment of VLMs in the context of these tasks. To be considered a correct prediction, the Euclidean distance between the predicted and ground truth box centers must be within a threshold, and the predicted category should also be accurate. Mathematically, this can be expressed as:

$$\text{Pr@k} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\Big(\|\hat{b}^i - b^i_{gt}\|_2 < k \bigcap (\hat{c}^i = c^i_{gt})\Big), \tag{3}$$

where N is the number of QA pairs, $\hat{b}^i$ and $\hat{c}^i$ denote the predictions for box center and object category corresponding to their annotation b^i_{gt} and c^i_{gt}, $\mathbb{1}$ is the indicator function, and k is the predefined distance threshold. We set Pr@1 as the primary metric in the following experiments.

Regarding the seven language-related tasks, we employ three established metrics, namely CIDEr [81], ROUGE-L [51], and BLEU [67]. In contrast to the simplistic word-wise evaluation of BLEU and ROUGE-L, CIDEr assesses sentences based on content and semantics, aligning more closely with human judgment [1]. Hence we employ CIDEr as the primary metric for evaluating the quality and correctness of output sentences. For the convenience of comparison, a rescaling involving $log_{10}(\text{CIDEr} + 1)$ is employed to standardize CIDEr values within the range of 0 to 1. In addition, we present the aggregate metric for BLEU by averaging values across BLEU-1 to BLEU-4.

Fig. 4. Visualization on the benchmark. We provide results for seven tasks through images and corresponding QA pairs. The remaining tasks are included in the Appendix.

Table 3. Comparison to State-of-the-arts. All methods are **fine-tuned** on the corresponding tasks. The main metrics (%) are marked in gray. **Bold** emphasizes top method; underline marks the runner-up. C: CIDEr; R: ROUGE-L; B: BLEU.

(a) nuScenes. ELM outperforms the leading previous methods on the matority of metrics across all six tasks on nuScenes, which validates the generality of our model.

Methods	Tracking		Box Detection		Box Prediction		Traffic Sign Inquiry			Surrounding Narration			Action & Decision		
	Pr@1	Pr@2	Pr@1	Pr@2	Pr@1	Pr@2	C	R	B	C	R	B	C	R	B
BLIP2-opt [48]	0.1	0.1	0.1	0.2	0.2	0.5	23.0	26.9	20.5	8.1	19.7	21.2	8.4	11.5	11.1
BLIP2-flant5 [48]	3.0	6.0	5.1	10.5	3.6	6.3	63.1	39.4	31.4	65.2	64.9	27.9	68.7	71.4	43.1
LLaMA-Ada. [27]	6.1	10.5	8.3	14.9	7.5	12.5	68.3	66.6	61.6	67.0	77.5	**60.1**	72.3	76.8	**64.7**
LLaVA [53]	5.5	9.3	28.5	31.2	6.1	10.2	51.1	58.5	50.8	64.9	64.6	41.2	64.4	62.4	57.9
Otter [46]	10.0	17.2	41.8	46.9	8.9	15.8	62.8	41.1	32.4	60.0	64.2	13.3	69.2	73.2	53.0
VideoChat [49]	0.4	0.9	0.1	0.3	0.1	0.2	25.3	21.9	11.7	21.7	29.2	12.2	29.6	33.2	13.1
Vid-ChatGPT [61]	0.1	0.6	0.1	1.0	0.3	1.2	49.6	57.1	48.6	61.0	69.6	37.2	53.6	58.5	43.5
ELM (Ours)	**14.0**	**23.3**	**51.6**	**56.9**	**15.1**	**24.4**	**76.5**	**71.2**	**63.9**	**73.2**	**78.7**	29.8	**74.4**	**83.3**	41.2

(b) Ego4D. We extend the model to Ego4D dataset and verified the generality of our token bank module on four tasks.

Methods	Moment Recap			Event Query			Ego. Narration			Activity Prediction		
	C	R	B	C	R	B	C	R	B	C	R	B
BLIP2-opt [48]	1.2	8.9	6.8	7.8	28.4	14.7	5.2	19.8	10.7	2.7	18.7	9.6
BLIP2-flant5 [48]	13.1	31.9	12.5	27.3	33.0	16.6	16.9	33.5	15.4	11.5	31.2	11.3
LLaMA-Ada. [27]	11.2	30.2	12.3	37.5	**47.2**	28.1	18.4	34.2	15.3	13.1	31.2	12.8
LLaVA [53]	9.6	28.3	12.1	**39.8**	44.6	**29.9**	6.5	28.2	11.6	8.4	28.0	13.0
Otter [46]	11.4	29.6	10.5	27.1	38.3	19.1	14.1	31.4	13.9	11.1	29.4	10.3
VideoChat [49]	13.2	32.5	13.8	34.5	42.2	26.4	20.7	35.0	**17.6**	12.1	32.4	14.1
Vid-ChatGPT [61]	10.0	31.1	13.3	27.9	36.5	20.9	10.2	21.7	10.4	9.4	30.5	12.6
ELM (Ours)	**22.6**	**36.7**	**19.4**	38.0	43.1	27.6	**26.5**	**37.7**	16.9	**18.1**	**34.1**	**17.0**

(c) Parameters.

Methods	Param.
BLIP2-opt	2.7B
BLIP2-flant5	2.7B
LLaMA-Ada.	7B
LLaVA	7B
Otter	7B
VideoChat	7B
Vid-ChatGPT	7B
ELM (Ours)	2.7B

4.1 Comparison to State-of-the-Arts

We first show the performance comparison of ELM and previous state-of-the-art VLMs [3,27,46,48,49,53,61] on our proposed benchmark. All VLMs are initialized using the official pre-trained weights and then fine-tuned on our dataset. Detailed metrics with respect to all tasks are documented in Table 3. On localization-related tasks such as Box Detection, our model attains a significant superiority. Notably, our method surpasses Otter [46] with a remarkable margin of **+9.8%** in Pr@1 score, illustrating the effectiveness of our proposed space-aware pre-training. On time-related tasks, *e.g.*, Traffic Sign Inquiry and Moment Recap, we surpass the second-best by **+13.4%** and **+6.8%** in CIDEr score, respectively. This highlights ELM's outstanding ability in retrieving timestamp information, attributed to the time-aware token selection. We notice that LLaVA [53] exhibits superior performance compared to ELM in Event Query task that focuses on successive event reasoning. ELM, which excels in precise timestamp retrieval, may face limitations in handling this specific task due to the inherent constraints in the FlanT5 [69] model's capacity to comprehend lengthy texts. Besides, due to the preference of our model for generating concise responses, its performance in terms of BLEU is affected [1]. Figure 4 demonstrates the qualitative comparison between ELM and baseline method (*i.e.*, BLIP2-flant5 [48]) on nuScenes [8] and Ego4D [28] dataset. It is observed that ELM's output is much closer to the ground truth, especially in tasks involving 3D localization. Additional visualizations are shown in the supplementary materials.

4.2 Ablation Study

We conduct ablation studies to assess the effectiveness of each component, with experiments shown in Table 4. Exp.0 serves as a baseline built upon BLIP2-

Table 4. Ablations on the effectiveness of each component. Baseline (Exp.0) uses the BLIP2-flant5 [48] model. ELM (Exp.7) is marked in gray. We only show the main metrics for brevity. `T`: Tracking, `BD`: Box Detection, `BP`: Box Prediction, `TSI`: Traffic Sign Inquiry; `SN`: Surrounding Narration; `AD`: Action & Decision; `EN`: Egocentric Narration; `MR`: Moment Recap; `EQ`: Event Query; `AP`: Activity Prediction; `Loc`: Localization; `Des`: Description; `C`: CIDEr; `R`: ROUGE-L; `B`: BLEU.

(a) Ablations on pre-training.

	Vocab	Data	T	BD	BP	TSI	SN	AD	EN
0	-	-	3.0	5.1	3.6	63.1	65.2	68.7	16.9
1	Rewritten	-	2.8	5.9	3.0	-	-	-	-
2	-	Loc	6.5	31.2	7.7	-	-	-	-
3	Rewritten	Loc	12.2	46.5	12.6	59.4	63.7	63.8	16.2
7	Rewritten	Des, Loc	**14.0**	**51.6**	**15.1**	**76.5**	**73.2**	**71.4**	**26.5**

(b) Ablations on token selection.

	Encoding	Selection	MR	EQ	AP
0	-	-	13.1	27.3	11.5
4	Sinusoidal	-	12.3	34.8	12.1
5	Textual	Hard	17.8	37.3	13.3
6	-	Manual	18.9	**39.4**	17.6
7	Textual	Soft	**22.6**	38.0	**18.1**

Table 5. Labeling quality and corresponding time cost. `Baseline`: LLaMA-Ada., $A_{\texttt{GPT}}$: accuracy between auto and manually annotated text evaluated by GPT, $S_{\texttt{GPT4V}}$: rationality score in image-text matching evaluated by GPT4V, $D_{\texttt{n-gram}}$: diversity evaluated by distinct n-gram ratio of phrases. Time refers to the average duration required for a single person to annotate a piece of data. Our choice is marked in gray.

Method	A_{GPT}	S_{GPT4V}	$D_{\text{n-gram}}$	Time(s/#)↓
Baseline	54.3	34.4	14.8	**1.6**
+ Filtering	68.3	49.5	21.2	1.9
+ Verification	**84.4**	**66.9**	**26.7**	4.5
Manual Labeling	*100*	*64.3*	*23.3*	*72.4*

Table 6. Planning on out-of-distribution datasets. Command Mean denotes the average value of the trajectories corresponding to each instruction in the training set. `ADE` & `FDE`: average & final distance error (m) of future trajectory in 3 s. All methods are trained on nuScenes [8] and evaluated on Waymo [78].

Method	ADE↓	FDE↓	Time(s)↓
Command Mean	7.98	11.41	-
UniAD-single [34]	4.16	9.31	0.56
Flamingo [3]	2.78	5.31	1.47
ELM	**2.28**	**4.27**	1.61

flant5 [48] and Exp.7 represents the final design of ELM. Initially, we examine the pre-training strategy within our pipeline in Table 4 (a). Comparative analysis between Exp.0 and Exp.2 reveals that solely performing localization pre-training without rewriting the space-relevant vocabulary yields limited improvements. Notably, the collaborative application of vocabulary rewriting and localization pre-training manifests a substantial advancement across all three localization tasks, exemplified by improvements of **+9.2%**, **+41.4%**, and **+9.6%** in Tracking, Box Detection, and Box Prediction, respectively. Nevertheless, a decrement in performance on alternative tasks is observed in Exp.3, prompting the adoption of cooperative pre-training in our final configuration (Exp.7), which brings enhanced performance across all tasks. This underscores the significance of integrating both localization and description data during the pre-training phase. We

note that the model's performance in the localization-related tasks improves by **+1.8%**, **+5.1%**, and **+2.5%** after the incorporation of descriptive data. We believe this is due to the fact that descriptive data provides information about relative positional relationships that benefits the localization tasks.

In addition, we explore several implementations of the token selection module, as results listed in Table 4 (b). The utilization of a straightforward sinusoidal temporal encoding may result in a marginal performance decline, potentially stemming from the model's difficulty in interpreting temporal information in this encoding scheme. It is worth noting that hard selection denotes selecting the tokens of three frames with the highest attention scores, while soft selection is the weighted summation across all tokens. Manual selection, which involves picking the tokens of the three frames based on the ground truth timestamp, is the theoretically optimal solution to hard selection. Using textual encoding strategy (as detailed in Sect. 3.3) with hard selection results in a noticeable improvement of **+4.7%**, **+10.0%**, and **+1.8%** on three tasks. Ultimately, we incorporate soft token selection, potentially encompassing information across all tokens, into our model. This adaptation brings improved performance on Moment Recap and Activity Prediction tasks, denoted as **+9.5%** and **+6.6%**, respectively, while preserving comparability with manual selection on the Event Query task.

Table 7. Extended evaluation on 3D detection performance. The Hungarian algorithm [44] is employed to ensure a reasonably fair comparison between ELM and conventional 3D detection models. `ped`: pedestrian; `bar`: barrier; `tra`: trailer. The main metric is marked in gray.

Method	$Pr^*@1$	$Pr^*_{car}@1$	$Pr^*_{ped}@1$	$Pr^*_{bar}@1$	$Pr^*_{tra}@1$
DETR3D [86]	43.6	48.9	44.6	39.1	15.6
BEVFormer [50]	47.4	52.3	48.8	43.5	14.3
VCD [35]	53.4	50.3	60.0	68.1	20.6
Method	Pr@1	$Pr_{car}@1$	$Pr_{ped}@1$	$Pr_{bar}@1$	$Pr_{tra}@1$
Ours	51.6	64.9	50.2	70.4	26.4

Table 8. Zero-shot evaluations on new tasks. Our model is also capable of achieving decent performance on zero-shot scenarios in comparison to supervised fine-tuning (SFT).

Task	Method	Pr@1	Pr@2	Pr@4
Tracking	SFT	**14.0**	**23.3**	**36.9**
	Zero-shot	9.8	14.8	23.0
Task	Method	C	R	B
Action & Decision	SFT	**71.4**	**74.6**	**43.0**
	Zero-shot	59.0	65.0	35.3

4.3 Further Discussions and Analysis

Evaluation on label quality and diversity. To verify the reliability of the auto-labeling pipeline, we manually annotate thousands of images and conduct a quantitative experiment for auto-labeling in Table 5. The baseline is an auto-labeling pipeline using LLaMA-Adater V2 [27]. A_{GPT} is the accuracy between automatically annotated text and human-annotated text evaluated by GPT4 [63], S_{GPT4V} is the rationality score in image-text matching evaluated

by GPT4V, and $D_{\text{n-gram}}$ is diversity, evaluated by different n-gram ratios in a phrase. Results show that our auto-labeling quality nearly equals manual annotation and leads in diversity (**26.7**), signifying the excellence of our data. Additionally, we report the average time required to collect each piece of data using different annotation methods. Manual image labeling entails meticulous inspection and detailed textual description, demanding significant time investment. Conversely, automated annotation pipelines enable annotators to efficiently filter and rectify errors in sampled image batches, substantially decreasing time consumption.

Out-of-Distribution Evaluation. To adequately demonstrate the model's generalization ability, Table 6 shows experiments on planning in unseen data, which includes both temporal and spatial understanding of ego vehicle future trajectory. Each frame in nuScenes is associated with one of 3 commands: `turn_left`, `turn_right`, or `go_straight`. The baseline, Command Mean, uses the mean of all trajectories in the training set whose command matches the current test frame command. Moreover, we compare our method with the current state-of-the-art method on nuScenes, UniAD [34]. In addition to the released checkpoint that requires multi-view input, we trained a single-frame version (UniAD-Single) for a fair comparison with our single-frame VLM. All methods are trained solely on the front view of nuScenes and are applied to Waymo without fine-tuning or adaptation. Please refer to the supplementary materials for the detailed design of using ELM for planning. ELM achieves respectable results in novel scenarios, surpassing end-to-end driving (UniAD) and other VLMs (Flamingo).

Comparison to Traditional 3D Perception Task. Addressing concerns pertaining to the superiority of embodied understanding over traditional 3D localization methods, our model is benchmarked against DETR3D [86], BEVFormer [50], and VCD [35], as listed in Table 7. Although our QA-based approach does not produce confidence scores, we have made efforts to conduct fair comparisons. The Pr*@1 metric is derived from (3) after performing a Hungarian algorithm [44] to establish a one-to-one matching between the prediction and the ground truth. The results show that ELM is comparable to classical models in 3D perception. Additional comparisons are in the supplementary materials.

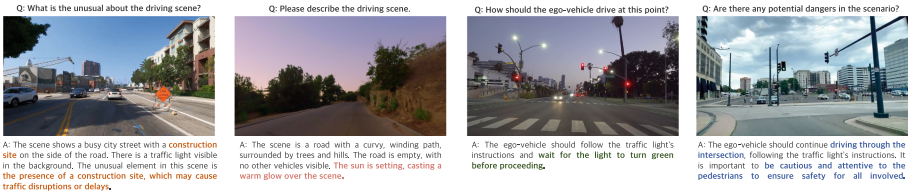

Fig. 5. Zero-shot on new scenarios. We select images from the internet that are not utilized during the training to assess the model's proficiency in unexplored scenarios. The results validate our model's ability to create notably logical interpretations.

Zero-Shot on New Tasks. To assess the generalization of ELM across different tasks within the benchmark, we fine-tune it using data associated with Box Detection and Moment Recap tasks, with subsequent testing on Tracking. Additionally, we fine-tune the model on Surrounding Narration and Activity Prediction, followed by inference on the Action & Decision task. The results in Table 8 indicate that the model's zero-shot capability, to handle tasks unseen before, is on par with supervised learning. Notably, even evaluated in a zero-shot manner, ELM performs comparably to the previous VLM on both tasks (see Table 3). We attain a zero-shot performance of **9.8%** in tracking compared to Otter's **10.0%**.

Verification of Open Scene Understanding. We evaluate ELM on novel scenarios and tasks to validate its generalization. The visualization in Fig. 5 demonstrates the superior scene understanding ability of our model on unseen data. Impressively, it can pay attention to construction signs on the road, make rational driving decisions, and analyze potential dangers. This showcases the potential to surpass traditional perception models in understanding unseen scenarios.

5 Conclusion and Limitation

We apply VLMs to achieve embodied understanding of driving scenarios and present a benchmark consisting of a suite of tasks and rubrics. ELM is proposed for the pursuit of understanding driving scenes in long-scope space and time, exhibiting promising generalization performance.

Limitations and Future Work. Currently, ELM only perceives driving scenes and interacts with human users. ELM can be further explored to generate driving control signals. Additionally, we will implement a prototype system, making ELM an embodied agent for closed-loop autonomous driving. Further experiments are needed to examine the model's capacity in broader scenarios, as our databases are mostly nuScenes [8] and Ego4D [28]. To promote the adoption of this model in real-world deployments, more validations need to be conducted to verify whether common sense reasoning helps decision-making in novel scenarios.

Acknowledgments. This work is supported by National Key R&D Program of China (2022ZD0160104, 2022YFB4501400), National Natural Science Foundation of China (62206172), and Shanghai Committee of Science and Technology (23YF1462000).

References

1. Aafaq, N., Mian, A., Liu, W., Gilani, S.Z., Shah, M.: Video description: a survey of methods, datasets, and evaluation metrics. ACM Comput. Surv. (CSUR) **52**(6), 1–37 (2019)
2. Abu-El-Haija, S., et al.: Youtube-8m: a large-scale video classification benchmark. arXiv preprint arXiv:1609.08675 (2016)

3. Alayrac, J.B., et al.: Flamingo: a visual language model for few-shot learning. Adv. Neural. Inf. Process. Syst. **35**, 23716–23736 (2022)
4. Brohan, A., et al.: RT-2: Vision-language-action models transfer web knowledge to robotic control. arXiv preprint arXiv:2307.15818 (2023)
5. Brohan, A., et al.: Rt-1: robotics transformer for real-world control at scale. arXiv preprint arXiv:2212.06817 (2022)
6. Brown, T.B., et al.: Language models are few-shot learners (2020)
7. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: a large-scale video benchmark for human activity understanding (2015)
8. Caesar, H., et al.: nuScenes: a multimodal dataset for autonomous driving (2020)
9. Casas, S., Sadat, A., Urtasun, R.: Mp3: a unified model to map, perceive, predict and plan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14403–14412 (2021)
10. Chen, G., et al.: Tem-adapter: adapting image-text pretraining for video question answer (2023)
11. Chen, L., et al.: Language models are visual reasoning coordinators. In: ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models (2023)
12. Chen, L., et al.: Driving with LLMS: fusing object-level vector modality for explainable autonomous driving. arXiv preprint arXiv:2310.01957 (2023)
13. Chu, X., et al.: Mobilevlm: a fast, reproducible and strong vision language assistant for mobile devices. arXiv preprint arXiv:2312.16886 (2023)
14. Chung, J.J.Y., Kamar, E., Amershi, S.: Increasing diversity while maintaining accuracy: text data generation with large language models and human interventions. arXiv preprint arXiv:2306.04140 (2023)
15. Dauner, D., Hallgarten, M., Geiger, A., Chitta, K.: Parting with misconceptions about learning-based vehicle motion planning. arXiv preprint arXiv:2306.07962 (2023)
16. Deruyttere, T., Grujicic, D., Blaschko, M.B., Moens, M.F.: Talk2Car: predicting physical trajectories for natural language commands. IEEE Access (2022)
17. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
18. Dewangan, V., et al.: Talk2BEV: language-enhanced bird's-eye view maps for autonomous driving. arXiv preprint arXiv:2310.02251 (2023)
19. Ding, X., Han, J., Xu, H., Zhang, W., Li, X.: HiLM-D: towards high-resolution understanding in multimodal large language models for autonomous driving. arXiv preprint arXiv:2309.05186 (2023)
20. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
21. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: CARLA: an open urban driving simulator. In: Proceedings of the 1st Annual Conference on Robot Learning, pp. 1–16 (2017)
22. Driess, D., et al.: PaLM-E: an embodied multimodal language model (2023)
23. Echterhoff, J., Yan, A., Han, K., Abdelraouf, A., Gupta, R., McAuley, J.: Driving through the concept gridlock: unraveling explainability bottlenecks. arXiv preprint arXiv:2310.16639 (2023)
24. Elhafsi, A., Sinha, R., Agia, C., Schmerling, E., Nesnas, I., Pavone, M.: Semantic anomaly detection with large language models (2023)
25. Fan, H., et al.: Baidu Apollo EM motion planner. arXiv preprint arXiv:1807.08048 (2018)

26. Fang, Y., et al.: Eva: exploring the limits of masked visual representation learning at scale. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19358–19369 (2023)
27. Gao, P., et al.: LLaMA-Adapter v2: parameter-efficient visual instruction model. arXiv preprint arXiv:2304.15010 (2023)
28. Grauman, K., et al.: Ego4d: around the world in 3,000 hours of egocentric video (2022)
29. Gu, J., et al.: Robotic task generalization via hindsight trajectory sketches. In: First Workshop on Out-of-Distribution Generalization in Robotics at CoRL 2023 (2023)
30. Hao, Y., et al.: Language models are general-purpose interfaces. arXiv preprint arXiv:2206.06336 (2022)
31. Hu, E.J., et al.: Lora: low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
32. Hu, P., Huang, A., Dolan, J., Held, D., Ramanan, D.: Safe local motion planning with self-supervised freespace forecasting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12732–12741 (2021)
33. Hu, S., Chen, L., Wu, P., Li, H., Yan, J., Tao, D.: St-p3: end-to-end vision-based autonomous driving via spatial-temporal feature learning (2022)
34. Hu, Y., et al.: Planning-oriented autonomous driving (2023)
35. Huang, L., et al.: Leveraging vision-centric multi-modal expertise for 3D object detection. arXiv preprint arXiv:2310.15670 (2023)
36. Huang, S., et al., et al.: Language is not all you need: aligning perception with language models. arXiv preprint arXiv:2302.14045 (2023)
37. Jin, B., et al.: Adapt: action-aware driving caption transformer (2023)
38. Karamcheti, S., et al.: Language-Driven representation learning for robotics (2023)
39. Kay, W., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
40. Keysan, A., et al.: Can you text what is happening? integrating pre-trained language encoders into trajectory prediction models for autonomous driving. arXiv preprint arXiv:2309.05282 (2023)
41. Khurana, T., Hu, P., Dave, A., Ziglar, J., Held, D., Ramanan, D.: Differentiable raycasting for self-supervised occupancy forecasting. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13698, pp. 353–369. Springer, Cham (2022)
42. Kim, J., Misu, T., Chen, Y.T., Tawari, A., Canny, J.: Grounding human-to-vehicle advice for self-driving vehicles (2019)
43. Kim, J., Rohrbach, A., Darrell, T., Canny, J., Akata, Z.: Textual explanations for self-driving vehicles (2018)
44. Kuhn, H.W.: The Hungarian method for the assignment problem. Naval Res. Logist. Quart. **2**(1–2), 83–97 (1955)
45. LeCun, Y.: A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. Open Review **62** (2022)
46. Li, B., et al.: MIMIC-IT: multi-modal in-context instruction tuning. arXiv preprint arXiv:2306.05425 (2023)
47. Li, H., et al.: Open-sourced data ecosystem in autonomous driving: the present and future (2023). https://doi.org/10.13140/RG.2.2.10945.74088
48. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models (2023)
49. Li, K., et al.: Videochat: chat-centric video understanding. arXiv preprint arXiv:2305.06355 (2023)

50. Li, Z., et al.: BEVFormer: learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13669, pp. 1–18. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-20077-9_1
51. Lin, C.Y.: Rouge: a package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81 (2004)
52. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
53. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2023)
54. Locatello, F., et al.: Object-centric learning with slot attention. Adv. Neural. Inf. Process. Syst. **33**, 11525–11538 (2020)
55. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
56. Lu, P., et al.: Learn to explain: multimodal reasoning via thought chains for science question answering (2022)
57. Majumdar, A., et al.: Where are we in the search for an artificial visual cortex for embodied intelligence? arXiv preprint arXiv:2303.18240 (2023)
58. Malla, S., Choi, C., Dwivedi, I., Choi, J.H., Li, J.: DRAMA: joint risk localization and captioning in driving (2023)
59. Mao, J., Qian, Y., Zhao, H., Wang, Y.: GPT-driver: learning to drive with GPT. arXiv preprint arXiv:2310.01415 (2023)
60. Mu, Y., et al.: Embodiedgpt: vision-language pre-training via embodied chain of thought. arXiv preprint arXiv:2305.15021 (2023)
61. Maaz, M., Rasheed, H., Khan, K., Khan, F.: Video-ChatGPT: towards detailed video understanding via large vision and language models. arXiv:2306.05424 (2023)
62. OpenAI, R.: Dall.e 3 system card (2023)
63. OpenAI, R.: GPT-4 technical report. arXiv pp. 2303–08774 (2023)
64. OpenAI, R.: GPT-4v(ision) system card (2023)
65. Padalkar, A., et al.: Open x-embodiment: Robotic learning datasets and rt-x models. arXiv preprint arXiv:2310.08864 (2023)
66. Palo, N.D., Byravan, A., Hasenclever, L., Wulfmeier, M., Heess, N., Riedmiller, M.: Towards a unified agent with foundation models. arXiv preprint arXiv:2307.09668 (2023)
67. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318 (2002)
68. Qian, T., Chen, J., Zhuo, L., Jiao, Y., Jiang, Y.G.: NuScenes-QA: a multi-modal visual question answering benchmark for autonomous driving scenario. arXiv preprint arXiv:2305.14836 (2023)
69. Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer (2020)
70. Regulation, G.D.P.: Art. 22 GDPR. automated individual decision-making, including profiling. Intersoft Consulting (2020)
71. Sachdeva, E., et al.: Rank2Tell: a multimodal driving dataset for joint importance ranking and reasoning. arXiv preprint arXiv:2309.06597 (2023)
72. Sauer, A., Savinov, N., Geiger, A.: Conditional affordance learning for driving in urban environments. In: Conference on Robot Learning, pp. 237–252. PMLR (2018)
73. Seff, A., et al.: MotionLM: multi-agent motion forecasting as language modeling (2023)

74. Sha, H., et al.: LanguageMPC: large language models as decision makers for autonomous driving. arXiv preprint arXiv:2310.03026 (2023)
75. Shah, D., et al.: VINT: a foundation model for visual navigation. arXiv preprint arXiv:2306.14846 (2023)
76. Sima, C., et al.: DriveLM: driving with graph visual question answering. arXiv preprint arXiv:2312.14150 (2023)
77. Song, E., et al.: MovieChat: from dense token to sparse memory for long video understanding. arXiv preprint arXiv:2307.16449 (2023)
78. Sun, P., et al.: Scalability in perception for autonomous driving: WAYMO open dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2446–2454 (2020)
79. Touvron, H., et al.: Llama 2: open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
80. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
81. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: consensus-based image description evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4566–4575 (2015)
82. Voigt, P., Von dem Bussche, A.: The EU general data protection regulation (GDPR). A Practical Guide, 1st Ed. **10**(3152676), 10–5555 (2017)
83. Wang, H., et al.: OpenLane-V2: A topology reasoning benchmark for unified 3D HD mapping (2023)
84. Wang, J., et al.: Git: a generative image-to-text transformer for vision and language. arXiv preprint arXiv:2205.14100 (2022)
85. Wang, P., Huang, X., Cheng, X., Zhou, D., Geng, Q., Yang, R.: The apolloscape open dataset for autonomous driving and its application. IEEE Trans. Pattern Anal. Mach. Intell. (2019)
86. Wang, Y., Guizilini, V.C., Zhang, T., Wang, Y., Zhao, H., Solomon, J.: Detr3D: 3D object detection from multi-view images via 3d-to-2d queries. In: Conference on Robot Learning, pp. 180–191. PMLR (2022)
87. Wayve: Lingo-1 (2023). https://wayve.ai/thinking/lingo-natural-language-autonomous-driving/
88. Wu, D., Han, W., Wang, T., Dong, X., Zhang, X., Shen, J.: Referring Multi-Object tracking (2023)
89. Wu, D., Han, W., Wang, T., Liu, Y., Zhang, X., Shen, J.: Language prompt for autonomous driving. arXiv preprint arXiv:2309.04379 (2023)
90. Xu, N., et al.: YouTube-VOS: a large-scale video object segmentation benchmark. arXiv preprint arXiv:1809.03327 (2018)
91. Xu, Y., et al.: Explainable object-induced action decision for autonomous vehicles (2020)
92. Xu, Z., et al.: DriveGPT4: interpretable end-to-end autonomous driving via large language model. arXiv preprint arXiv:2310.01412 (2023)
93. Yang, Z., Jia, X., Li, H., Yan, J.: A survey of large language models for autonomous driving (2023)
94. Zeng, W., et al.: End-to-end interpretable neural motion planner. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8660–8669 (2019)
95. Zhai, Y., et al.: Investigating the catastrophic forgetting in multimodal large language models. arXiv preprint arXiv:2309.10313 (2023)
96. Zhang, P., Zeng, G., Wang, T., Lu, W.: Tinyllama: an open-source small language model. arXiv preprint arXiv:2401.02385 (2024)

97. Zhang, Q., Peng, Z., Zhou, B.: Learning to drive by watching YouTube videos: Action-conditioned contrastive policy pretraining. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13686, pp. 111–128. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19809-0_7
98. Zhang, R., et al.: LLaMA-adapter: efficient fine-tuning of language models with zero-init attention. arXiv preprint arXiv:2303.16199 (2023)
99. Zhu, H., et al.: CelebV-HQ: a large-scale video facial attributes dataset. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13667, pp. 650–667. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-20071-7_38