Vision-Language Modeling for Scene Understanding and Reasoning of Vehicle-to-X Interactions

Hao Wang Suining He
{hao.3.wang, suining.he}@uconn.edu
Ubiquitous & Urban Computing Lab, School of Computing
University of Connecticut

***Abstract*—In the complex traffic environments, understanding how a focal vehicle interacts (e.g., maneuvers) with various traffic elements (e.g., other vehicles, pedestrians, and road infrastructures), i.e., vehicle-to-X interactions (VXIs), is essential for developing the advanced driving support and intelligent vehicles. To derive the VXI scene understanding, reasoning, and decision support (e.g., suggesting cautious move in response of a pedestrian crossing the street), this work takes into account the recent advances of multi-modality large language models (MLLMs). We develop VXI-SUR, a novel VXI Scene Understanding and Reasoning system based on vision-language modeling. VXI-SUR takes in the visual VXI scene, and generates the structured textual responses that interpret the VXI scene and suggests an appropriate decision (e.g., braking, slowing down). We have designed within VXI-SUR a VXI memory mechanism with both scene and knowledge augmentation mechanisms, and enabled scene-knowledge co-learning to capture complex correspondences across scenes and decisions. We have performed extensive and comprehensive evaluations of VXI-SUR based on an open-source dataset with ∼17k VXI scenes. We have conducted extensive experimentation studies upon VXI-SUR, and corroborated VXI awareness, description preciseness, semantic matching, and quality in understanding and reasoning the complex VXI scenes.**

I. Introduction

Understanding and reasoning a focal vehicle's interaction scene with various traffic elements, such as other vehicles, pedestrians, and road infrastructures (e.g., traffic lights), is essential for enhancing situational awareness and decision-making process of the advanced driving assistance systems (ADASs) and emerging connected autonomous vehicles (CAVs).

By interpreting such Vehicle-to-X Interactions (VXIs), the vehicles expresses awareness of the critical interactions in human-understandable natural language, such as the appearance of pedestrians crossing the street or sharing the road with a cyclist, and explains such scenes to the drivers or riders. The human drivers can, for instance, take responsible or defensive driving strategies [1], [2], and the CAV riders can gain improved situational awareness (or perform timely take-over actions if necessary). These will help improve the usefulness, trustworthiness, and adoption rates of emerging advanced mobility systems [3], [4].

Our studies here aim to develop a novel VXI scene understanding and reasoning system. In particular, we take into account the recent advances in the Multi-modality Large Language Models (MLLMs) [5]–[9]. Via MLLMs, our goal is to enable the understanding and reasoning capabilities in interpreting the interaction scenes and the critical objects, and generating human-understandable or explainable decision support (e.g., suggestion of slowing down, or braking [10]). Specifically, in this prototype study, we focus on visual (e.g., images) and textual modalities [9], and design a **VXI Scene Understanding and Reasoning** system based on **V**ision-**L**anguage **M**odeling, namely **VXI-SUR**.

Toward development of VXI-SUR, we particularly focus on addressing the following two important research challenges:

(A) How to gain semantic correspondences across scenes and decisions in complex VXIs? The complex VXI scenes may often be associated with various potential traffic conditions or contexts, and lead to a variety of possible decisions. As revealed in our data-driven studies, even the very similar VXI scenes, such as the appearance of a pedestrian, may correspond to different reasonable decisions, such as braking or maneuvering, due to various traffic environments, contexts, and factors (e.g., spatial distributions of other objects, width of the road). Gaining VXI understanding and reasoning, therefore, hinges on capturing such complex and long-tail correspondences across scenes and decisions. Despite the prior MLLM studies on the autonomous driving [11], how to strengthen the MLLM's capabilities in understanding and reasoning the semantic scene-to-decision correspondences remains largely under-explored.

(B) How to gain holistic and structured VXI reasoning beyond general scene interpretation? In developing such a VXI-SUR, gaining *holistic* VXI reasoning requires injection of structured formation of VXI scene data, from the visual knowledge of the scene to various important interaction-related factors (e.g., the concerned objects of interest and their types or categories, their relative closeness to the focal vehicle). Without a properly structured VXI data formation and well-defined MLLM fine-tuning (i.e., training the MLLM backbone to adapt to a specific domain) pipeline, the resulting VXI-SUR will lack sufficient knowledge in the VXI domain (say, how to respond when sharing the road with an incoming cyclist) in reasoning the scenes, and degrade the system usefulness in supporting ADAS or CAV users [11].

To overcome the above technical challenges, we have made the following technical contributions toward designs of VXI-SUR:

(1) Scene-knowledge augmentation mechanism for a VXI memory. We have designed a VXI memory based on a scene-knowledge augmentation mechanism. Such a

VXI memory enables VXI-SUR in capturing the complex correspondences among VXI scenes and decisions, and establishing the structured and holistic knowledge regarding the VXI scenes. In structuring VXI scene data for the MLLM fine-tuning, for each target VXI scene to be learned, VXI-SUR retrieves several similar VXI scenes to augment its memory regarding the complex scene-decision correspondences. Furthermore, for each VXI scene to be learned, VXI-SUR extracts important knowledge of the objects of interest (OoIs) in VXI, and their relative positions and closeness to the focal vehicles, to augment its memory with these contextual and environmental factors. Via our scene-knowledge augmentation designs, VXI-SUR forms the VXI memory, which helps adapt the MLLM backbone in understanding and reasoning the complex VXIs.

(2) VXI scene-knowledge co-learning. Based on the VXI memory mechanism, we have designed a novel scene-knowledge co-learning mechanism. Such a mechanism provides operations of (i) VXI template integration and (ii) gated scene fusion, which respectively fuse the textual and visual modalities. Specifically, we have designed the VXI template as the prompt to incorporate structured information regarding the VXI scene to the MLLM. The VXI scenes to be learned, along with their associated similar scenes, decisions, and auxiliary VXI knowledge, are *jointly* injected into the fine-tuning process of a general-purpose MLLM backbone. Gated scene fusion adapts the importance of visual modalities from the VXI memory. The VXI scene-knowledge co-learning yields holistic VXI scene understanding and reasoning (e.g., suggestion of autonomous vehicle control or vehicle maneuvering).

(3) Extensive VXI data studies and multifaceted experimental evaluations. We have conducted extensive VXI studies based on a real-world open-source dataset named DRAMA [12] (*D*riving *R*isk *A*ssessment *M*echanism with *A* captioning module from the Honda Research Institute). We have experimented our VXI-SUR based on a total of 16,938 VXI scenes, and fine-tuned a vision-language model backbone (InternVL [9] in our current study). We have conducted comprehensive studies to evaluate VXI awareness (i.e., accuracy in recognizing VXI types), description precision and semantic matching (i.e., how the generated scene descriptions match annotations), and quality of the generated responses (i.e., completeness of semantics based on GPT scores). Our extensive experimental results have demonstrated effectiveness and usefulness of VXI-SUR in understanding (e.g., describing the scenes) and reasoning (e.g., providing decision support in responses) the VXI scenes compared with the baselines.

II. Related Work

With the recent advances of language models, a myriad of multi-modality large language models (MLLMs) [9], [13]–[17], have emerged to support general-purpose multi-modality (e.g., vision and language) understanding and reasoning. Driven by these endeavors, the MLLM designs have been recently considered for the interaction scenarios of autonomous driving, where the MLLMs can serve as the engines to perceive, predict, and perform the decision-making in the complex traffic environments. For instance, Talk2BEV [18] accounted for the bird-eyes-view maps, and catered for different tasks in the autonomous driving, and similar studies have been conducted by BEV-InMLLM [19] to expand the situational awareness. CoDrivingLLM [20] studied an interactive and learnable cooperative driving and leveraged retrieval augmented generation to avoid repeating driving mistakes. Pittawat et al. [21] proposed large language model distillation for faster inference in complex driving scenes.

MLLMs have also been considered for expanding the interaction and communication capabilities across the mobility systems and their users [22], [23]. For instance, DriVLMe [22] investigated the natural and effective communication between vehicles and human riders in order to accommodate the dynamic traffic environments and task changes. Similarly, LanguageMPC [23] took into account the MLLMs to generate decisions in human-understandable free-form texts, and then translate them into actionable driving controls. However, these studies have not yet considered how to leverage the MLLMs to interpret and reason the VXI scene and communicate its decision support for the mobility system users.

III. System Framework Overview

We illustrate the entire system framework of VXI-SUR in Fig. 1. VXI-SUR is composed of two phases: model fine-tuning and model inference.

Specifically, in (I) the model fine-tuning phase, VXI-SUR conducts the operations of (a) VXI data preparation; (b) scene-knowledge augmentation; and (c) scene-knowledge co-learning. Then, in (II) the model inference phase, VXI-SUR performs the operation of (d) VXI scene understanding and reasoning. We note that our VXI-SUR framework, while focusing on vision-language modeling in this prototype study, is general enough to be extended to other MLLM backbones [9], [16] and sensing modalities (e.g., LiDAR).

(a) VXI Data Preparation. In this operation, VXI-SUR processes the VXI scenes (e.g., visual modality), and important attributes in terms of textual modality, i.e., VXI type (for instance, vehicle-to-vehicle or vehicle-to-pedestrian interactions), VXI location description (the location of the traffic element or road infrastructure relative to the focal vehicle), scene description (i.e., the textual explanation of the VXI scene), and decision type (categories of different decisions).

Then, VXI-SUR structures the VXI scene data (visual and textual modalities in this study) based on a template generator, and structure them into the VXI textual templates. Each VXI textual template consists of the roles and tasks as inputs to the MLLM backbone (in the form of textual prompts), and the expected responses (outputs from the MLLM backbone). The VXI textual templates are fed to (b) VXI Scene-Knowledge

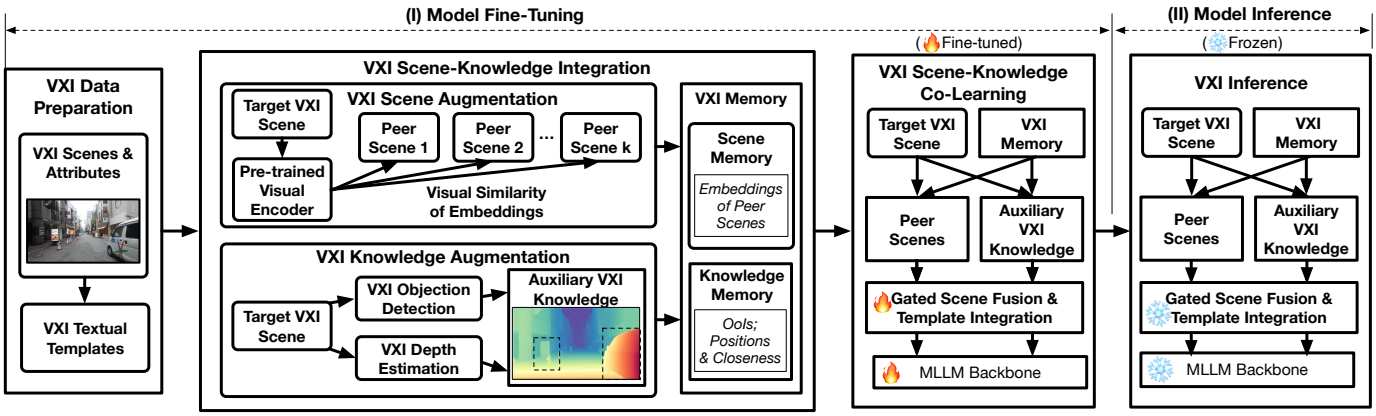

Figure 1: Overview of the framework of VXI-SUR, which consists of (I) model fine-tuning and (II) model inference. VXI-SUR performs the operations of (a) VXI data preparation, (b) VXI scene-knowledge augmentation, and (c) VXI understanding and reasoning inference.

Augmentation operation for VXI memory and subsequent MLLM fine-tuning.

(b) VXI Scene-Knowledge Augmentation. This operation aims to gain the VXI awareness and strengthen knowledge regarding the complex VXI scenes. We have designed two augmentation methods to retrieve the peer scenes and auxiliary VXI knowledge, and construct the VXI memory.

Specifically, for each target VXI scene, the VXI scene augmentation operation retrieves multiple peer scenes that are similar to the the target VXI scene based on visual similarities. In the meantime, for each VXI scene, the knowledge augmentation operation extracts all the objects of interest (OoIs) in the target VXI scene, i.e., the traffic elements (e.g., pedestrians, vehicles) and road infrastructures, and their position and closeness (depth) information based on the pre-trained VXI object detection and depth estimation models. The peer scenes, and auxiliary VXI knowledge of the target VXI scenes will form the VXI memory for the subsequent VXI scene-knowledge co-learning.

(c) VXI Scene-Knowledge Co-Learning. This operation aims to fine-tune the MLLM backbone. For each target VXI scene in the batch, VXI-SUR incorporates its peer scenes and the auxiliary VXI knowledge from the VXI memory for the MLLM fine-tuning. VXI-SUR performs both the VXI template integration and gated scene fusion within the co-learning process, and fine-tunes the model parameters of the MLLM backbone (in this prototype study we adopt open-sourced InternVL [9] as the backbone).

(d) VXI Inference. Given the fine-tuned MLLM backbone, VXI-SUR takes in the visual modality (say, the image) of a target VXI scene, and performs the VXI scene-knowledge augmentation by finding its peer scenes and auxiliary VXI knowledge. Based on the retrieved peer scenes and auxiliary VXI knowledge, the fine-tuned MLLM backbone (frozen) makes the online VXI inference, and generates responses of the VXI scene understanding and reasoning. Specifically, VXI-SUR generates the VXI responses (in texts) that include (i) the VXI scene understanding and reasoning text (i.e., a free-form and human-understandable textual explanation of the VXI scenes and decision suggestion) as well as (ii) the VXI scene attributes, i.e., VXI type, decision type, VXI location description, and scene description.

IV. VXI Data Preparation

A. VXI Scene Data Preprocessing

Toward development of VXI-SUR, we have taken into account the open-sourced real-world DRAMA (*D*riving *R*isk *A*ssessment *M*echanism with *A* captioning module) dataset [12] provided by the Honda Research Institute (HRI), US [24]. This dataset provides various VXI scenes, each of which consists of an image of the VXI scene, as well as attributes of VXI type, VXI location description, scene description, and decision type. In preprocessing the DRAMA dataset, we have filtered out the VXI scenes that contain no VXI type or scene description and obtained a total of 16,938 VXI scenes for our VXI-SUR development.

When processing each VXI scene, we have retrieved from the DRAMA dataset the following four VXI scene attributes in textual modalities: (i) the VXI type (i.e., type of OoIs interacting with the focal vehicle) in four categories, i.e., “vehicles”, “pedestrians”, “cyclists”, and “infrastructures”; (ii) VXI location description relative to the focal vehicle (e.g., “in front of the ego car” or “to the left side of the intersection”); (iii) scene description (for instance, “there is a black sedan slowing in the ego lane, in front of the ego car, because of traffic congestion ahead”); and (iv) decision type (i.e., type of actions that the focal vehicle should take) in eight categories — that is, “stop”, “follow”, “slow down”, “be cautious”, “carefully maneuver”, “yield”, “start moving”, and “accelerate”. We note that in terms of road infrastructures, we take into account the common objects such as traffic lights, stop signs, and cones.

In terms of VXI types, we note that we have identified 12,212, 2,853, 431, and 1,441 VXI scenes that pertain to interactions with (a) vehicles, (b) pedestrians, (c) cyclists, and (d) infrastructures, respectively.

System Prompt

- Your are a driving assistant, sitting in the copolit position of the car.
- You are supposed to notice the potential risk and give reasonable suggestions.
- You will be given some attributes and your job is to provide the suggestion and explanation for the driver.
- Think step by step and keep following the traffic laws and moral codes.
- Your suggestions should limit to the given information.
- Your output should not use new lines.

User Prompt

- Description of the scene is: **{scene description},**
- Involved object of interest is **{VXI Type},**
- Location is **{VXI location description},**
- Suggestion is **{decision type},**

(a)

Roles and Tasks for MLLM backbone

- You are a driving assistant.
- Your job is to find potential risk in the current driving scene.
- Please describe the potential risk in the **<image>** as well as the suggenstion for the driver.
- Your output should be two parts: attributes level and reasoning level.
- Your output should be json format.

Expected Responses

{
"Scene Description": "There is a black hatchback slowing down in the ego lane, in front of the ego car, because there is traffic congestion",
"VXI Type": "Vehicle",
"Decision Type": "Slow down",
"VXI Location Description": "ego lane",
"Scene Understanding and Reasoning Text": "I suggest that you slow down. The black hatchback in front of us is decelerating due to traffic congestion, which means we need to maintain a safe following distance to avoid a collision. By reducing our speed, we can better react to the changes in traffic and ensure the safety of everyone on the road. Remember, it's important to stay calm and patient in congested conditions.",
}

(b)

Figure 2: Examples of (a) a template generation prompt and (b) a VXI textual template. In (a) the template generation prompt, the texts highlighted in red will embed the needed text from the VXI scene dataset. In (b), the "scene understanding and reasoning text" will be generated from (a), and the other fields are retrieved from the original VXI scene dataset.

B. VXI Textual Template Generation

To adapt the MLLM backbone for VXI understanding and reasoning, we have designed a VXI textual template that provides holistic and structured characterization of the target VXI scene to be learned. The template provides the understanding and reasoning texts regarding the VXI scene as well as the VXI scene attributes. We aim to leverage such a VXI textual template to help VXI-SUR gain a holistic, structured, and comprehensive view regarding the complex VXI scenes, and hence the MLLM can capture the important information for adapting to the complex VXI scene understanding and reasoning task.

To prepare such a VXI textual template, we leverage a large language model (LLM) as a template generator, such that a free-form and human-readable text pertaining to VXI scene understanding and reasoning can be provided. In this prototype study, we use GPT-3.5 [25] as the template generator for structuring the text into the JSON format. We have empirically evaluated and validated the performance of our template generator. We leverage the template generation prompt for the LLM, which consists of (i) the system prompt that includes the basic requirements in VXI textual template, and (ii) the user prompt that includes the four scene attributes, i.e., VXI type, decision type, VXI location description, and decision suggestion in the textual modalities.

We illustrate an example of the VXI template generation prompt in Fig. 2(a), which consists of the system prompt (top) and the user prompt (bottom). The system prompt in our VXI template serves as the requirement to constrain the structure and format of the generated template, while the user prompt serves as the variable based on which the template generator transforms into the final human-readable understanding and reasoning text.

Specifically, in terms of the system prompt, we have provided a set of prompt commands in order to generate the expected VXI textual template. These include (i) the role that the template generator needs to play; (ii) the definition of generation task, inputs, and expected outputs; (iii) guidance and steps that the generator needs to follow [26]; and (iv) constraints regarding the output content and format. The user prompt allows us (as the user of the template generator) to embed the needed texts in terms of the four scene attributes (scene description, VXI type, VXI location description, and decision type).

Based on the template generation prompt, the template generator transforms the four attributes into the VXI textual template. We illustrate a VXI textual template in Fig. 2(b), which consists of two components: (i) roles and tasks for the MLLM backbone (with an image placeholder, `<image>`, that points to the embedded visual modality); and (ii) expected responses (in JSON format for ease of parsing) from the MLLM backbone, which includes the four scene attributes as well as the understanding and reasoning text. The roles and tasks are fed as the assistant prompt to fine-tune the MLLM backbone, steering the MLLM to generate the texts fulfilling the expected responses.

V. Scene-Knowledge Augmentation & Co-Learning

Fig. 3 overviews the VXI scene and knowledge augmentation of a VXI scene example. For each target VXI scene, VXI-SUR augments it with several similar peer scenes, and integration of auxiliary VXI knowledge such as the OoIs and the position/depth information.

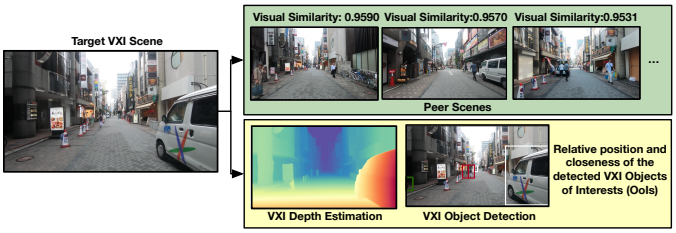

Figure 3: Overview of our proposed scene and knowledge augmentation.

A. VXI Scene Augmentation

The key idea of our VXI scene augmentation is to strengthen the awareness of semantic correspondence across VXI scenes

and decisions, by associating a target VXI scene with visually similar peer scenes as well as their important attributes, i.e., scene descriptions, decision types, and understanding and reasoning texts. This way, VXI-SUR can strengthen the learnability of its MLLM backbone and mitigate the impacts of the heterogeneous and long-tail nature of the VXIs.

Specifically, we first pass the images of the VXI scenes through a pre-trained visual encoder to obtain the embeddings. In this study, we adopt the pre-trained visual encoder of CLIP (Contrastive Language-Image Pre-training) [27], denoted as $\text{Enc}^{(\text{CLIP})}(\cdot)$, as it characterizes and captures the correlations of the input images [27]. Given the image of the target VXI scene, we have the embeddings as

$$\mathbf{E}_i^{(\text{CLIP})} = \text{Enc}^{(\text{CLIP})}(\mathbf{I}_i) \in \mathbb{R}^{512}. \quad (1)$$

Then, for the embedding $\mathbf{E}_i^{(\text{CLIP})}$ of each target VXI scene i, we find among the other scenes $\mathbf{E}_j^{(\text{CLIP})}$'s ($j \neq i$) the top k most similar embeddings, based on cosine similarity, i.e.,

$$\text{sim}\left(\mathbf{E}_i^{(\text{CLIP})}, \mathbf{E}_j^{(\text{CLIP})}\right) = \frac{\mathbf{E}_i^{(\text{CLIP})} \cdot \mathbf{E}_j^{(\text{CLIP})}}{||\mathbf{E}_i^{(\text{CLIP})}|| \cdot ||\mathbf{E}_j^{(\text{CLIP})}||}, \quad (2)$$

as the peer scenes to form the scene memory. The similarity calculation and scene memory retrieval can be further facilitated through the vector database [28], [29], which will be considered in our future studies.

We further illustrate a target VXI scene and three of its most similar peer scenes in Fig. 3. We can observe the similar traffic environments from the peer scenes, despite different VXI types and decision types (from left to right: "be cautious", "stop", and "slow down"). Based on the VXI scene augmentation, our subsequent scene-knowledge co-learning captures the subtle differences across the target VXI scene and its peer scenes, and therefore provides the more reasonable VXI decision support and mitigates the impacts of heterogeneous and long-tail correspondences.

B. VXI Knowledge Augmentation

While the peer VXI scenes and their attributes provide augmented information of scene-decision correspondences, they cannot provide the structured and spatial knowledge regarding the target VXI scene. Therefore, we further consider the VXI knowledge augmentation, and in this prototype study, we take into account identification of the OoIs and estimation of their relative positions and depths (closeness to the focal vehicle).

Specifically, we leverage the pre-trained VXI object detection and depth estimation models, Yolov8 [30] and DepthAnythingV2 [31] in our prototype study, to respectively retrieve the auxiliary knowledge on the involved OoIs and their rectangular bounding boxes (each is represented by its four coordinates). Each OoI is associated with the average depth values (estimated from DepthAnythingV2 [31]) over the pixels within its bounding box. In our current studies, the identified OoIs involve: (i) vehicles (such as 52,537 cars, 19,278 trucks, 4,643 buses, 2,848 bicycles, and 1,103 motorcycles), (ii) pedestrians (34,088 people), and (iii) infrastructures (such as 4,944 traffic lights, 332 stop signs, 152 fire hydrants, 21 parking meters) from all of our VXI scenes. These OoIs and their depth values come form the knowledge memory.

As illustrated in the example in Fig. 3, we obtain the bounding boxes of the identified OoIs (e.g., vehicle, pedestrian, and bicycles), and we also show the corresponding heatmap of the depth estimations, where the warmer colors (transformed from the grayscale depth image) represent the smaller depth values and closer objects.

Given above, we incorporate the VXI knowledge into our VXI textual template as follows. Specifically, as illustrated in Fig. 4, for each identified OoI, based on the bounding box estimated from the pre-trained Yolov8, we retrieve the horizontal coordinate of the center point of the OoI. We horizontally split the VXI image into three rectangular regions of equal height and width. We check the region that the center of the OoI falls in and annotate the label of "left", "center", or "right". In the meantime, based on the bounding box of the OoI, we find the relative closeness of the OoI based on average depth values of all the pixels within its bounding box. Based on the minimum and maximum depth values inside a VXI image, we project the relative closeness of a OoI into the range of [0, 255], and annotate the relative closeness based on the intervals of "far", "medium", or "close" (in this study, we use the intervals of [0, 50], (50, 150], and (150, 255]).

In the example of our target VXI scene (Fig. 4), the horizontal coordinate center of the OoI's bounding box (car) versus the image width is 0.85, while the average depth within the bounding box is 165. Therefore, we will have [relative position, closeness] of the OoI as [right, close].

Through these two transformations, we obtain the position knowledge of each OoI for our knowledge augmentation. As illustrated in Fig. 4, we derive the text block of auxiliary VXI knowledge, which summarizes the OoIs within the VXI scene (e.g., "there is 1 car in the VXI scene"), and describes the OoIs' position and closeness relative to the focal vehicle (e.g., "a car is to the right of the focal vehicle, and its distance is close"). This text block will serve as part of the inputs for the subsequent scene-knowledge co-learning (VXI template integration in Sec. V-C).

C. Scene-Knowledge Co-Learning

We have illustrated the designs of scene-knowledge co-learning in Fig. 5, which consists of (a) VXI template integration, and (b) gated scene fusion.

(a) VXI Template Integration. We further integrate the VXI scenes and knowledge into our textual input for the scene-knowledge co-learning. Our textual input consists of three parts, (i) roles and tasks for MLLM backbone from our VXI textual template (Fig. 2); (ii) attributes of peer scenes; and (iii) auxiliary VXI knowledge. This way, we realize the scene-knowledge augmentation upon the VXI template, which serves as the augmented textual input for co-learning.

Specifically, in terms of scenes, for each target VXI scene, we will include the important attributes of all the k peer scenes, i.e., decision types, scene descriptions, and VXI scene

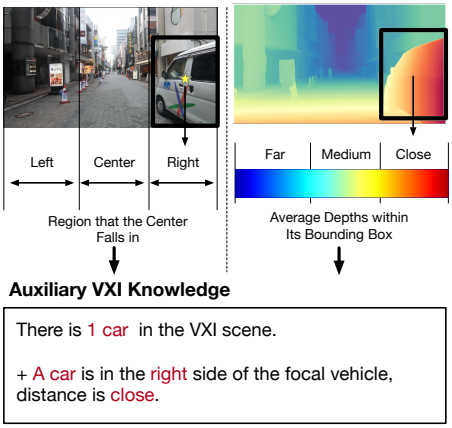

Figure 4: Illustration of our auxiliary VXI knowledge derivation in a target VXI scene.

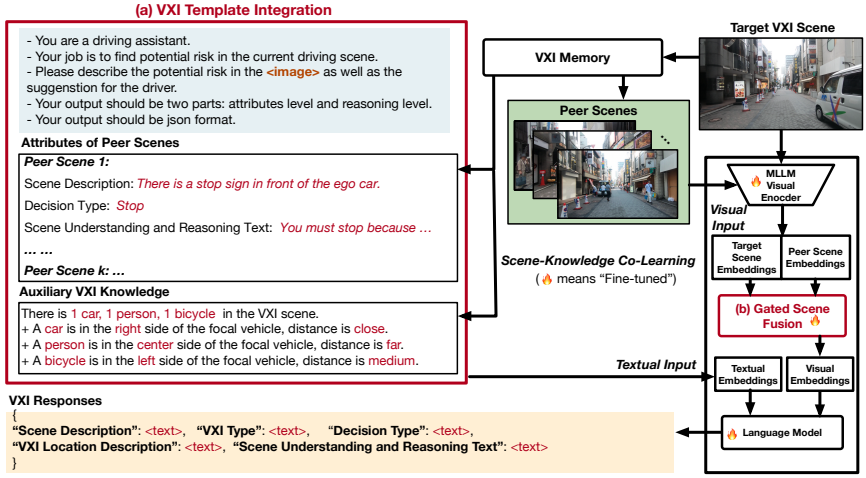

Figure 5: Overview of our scene-knowledge co-learning designs for MLLM fine-tuning, which consists of the two major designs: (a) VXI template integration and (b) gated scene fusion.

understanding and reasoning texts. In terms of knowledge, we will first summarize the detected OoIs, and integrate the texts of auxiliary VXI knowledge regarding the position and closeness knowledge of each OoI (as shown in Fig. 4). As illustrated in Fig. 5, these two blocks of texts are integrated with the roles and tasks for the MLLM backbone, and form the textual input for the subsequent scene-knowledge co-learning.

(b) Gated Scene Fusion. We have designed the gated scene fusion to incorporate visual modalities (images) from the peer VXI scenes, such that VXI-SUR can interpret the scenes from the VXI memory and reason about their differences and subsequent decisions. We note that our gated scene fusion leverages the visual encoder component available in the MLLM backbones, and hence our scene-knowledge co-learning designs can be easily applied to various vision-language model frameworks [9], [16].

Specifically, let $\text{Enc}^{(\text{MLLM})}(\cdot)$ be the MLLM visual encoder of the MLLM backbone (e.g., InternViT of InternVL in our case [9]). For each target VXI scene, we feed its image $\mathbf{I}$ through the visual encoder, and obtain

$$\mathbf{E}^{(\text{target})} = \text{Enc}^{(\text{MLLM})}(\mathbf{I}), \tag{3}$$

where $\mathbf{E}^{(\text{target})} \in \mathbb{R}^{m\times d}$, m represents number of visual tokens, and d represents the dimension of visual embeddings. In the meantime, given the k images of the peer VXI scenes retrieved from the VXI memory, i.e., $\{\mathbf{I}_1, \mathbf{I}_2, \ldots \mathbf{I}_k\}$, we feed each one of them through the MLLM visual encoder $\text{Enc}^{(\text{MLLM})}(\cdot)$, i.e.,

$$\mathbf{E}_i^{(\text{peer})} = \text{Enc}^{(\text{MLLM})}(\mathbf{I}_i), \tag{4}$$

where $\mathbf{E}_i^{(\text{peer})} \in \mathbb{R}^{m\times d}$. We further concatenate all the visual embeddings along their last dimension, i.e.,

$$\mathbf{E}^{(\text{peer})} = \texttt{concat}\left(\left[\mathbf{E}_1^{(\text{peer})}, \mathbf{E}_2^{(\text{peer})}, \ldots, \mathbf{E}_k^{(\text{peer})}\right]\right), \tag{5}$$

where the resulting embeddings $\mathbf{E}^{(\text{peer})} \in \mathbb{R}^{m\times(dk)}$.

Given above, we find the gated weight values, i.e.,

$$\mathbf{G} = \sigma\left(\mathbf{E}^{(\text{peer})} \cdot \mathbf{W}_G + b_G\right) \in \mathbb{R}^{m\times k}, \tag{6}$$

where σ represents the sigmoid activation function, $\mathbf{W}_G \in \mathbb{R}^{(dk)\times k}$ and b_G are learnable parameters. The resulting final visual embeddings are therefore given by

$$\mathbf{E}^{(\text{visual})} = \mathbf{E}^{(\text{target})} + \sum_{i=1}^{k} \mathbf{G}[:, i] \otimes \mathbf{E}_i^{(\text{peer})}, \tag{7}$$

where $\mathbf{E}^{(\text{visual})} \in \mathbb{R}^{m\times d}$, and $\otimes$ represents the Hadamard element-wise product. The final visual embeddings will be further fed to the MLLM backbone along with the corresponding VXI textual template (transformed into textual embeddings by text embedding layers within the MLLM backbone) as the inputs for fine-tuning.

Specifically, let $\mathbf{E}^{(\text{text})} \in \mathbb{R}^{n\times d}$ be the textual embeddings from the VXI textual template (n is the number of textual tokens), and $\left[\mathbf{E}^{(\text{visual})}, \mathbf{E}^{(\text{text})}\right] \in \mathbb{R}^{(m+n)\times d}$ be the combined input to the language model (e.g., the language part of InternVL in our case [9]). We fine-tune the language model by minimizing the auto-regressive loss, the average cross-entropy of predicting next token based on preceding ones [32].

VI. Experimental Evaluations

A. Experimental Settings

• **Baseline Approaches.** We compare the performance of VXI-SUR with the following baseline and state-of-the-art (SOTA) approaches: (1) ChatGPT-4o-Mini [5], (2) Gemini-Flash [6], (3) Claude-Haiku [7], (4) DeepSeek-VL [8] (a chat model with 1.3B parameters), and (5) InternVL [9] (a chat model with 1.0B parameters; with LoRA [33] fine-tuning enabled with a dimension of 128). We also compare VXI-SUR against its variations of (6) VXI-SUR w/o scene augmentation, (7) VXI-SUR w/o knowledge augmentation, and (8) VXI-SUR w/o VXI memory. For the online baselines (1)–(3), we follow their corresponding API requirements and transform their outputs into JSON format. For the local baselines (4) and (5), we follow the model and hyperparameter settings from their best practices provided in model fine-tuning benchmarking [34].

• **Evaluation Metrics.** In order to comprehensively evaluate performance of VXI-SUR compared with other baseline approaches, we have designed the following three sets of metrics.

- **VXI Awareness**: We quantify the VXI awareness based on the multi-class classification accuracy regarding the estimated categories of VXI types. The higher accuracy in the classification of VXI types implies the more effectiveness in understanding the VXI scenes.
- **Preciseness**: We quantify the preciseness of the scene description based on conventional metrics of BLEU-1 [35], METEOR [36], Rouge-L, and Rouge-1 [37]. The higher values in these metrics indicate better preciseness and matching of generated scene descriptions with the annotated (ground-truth) ones.
- **Semantic Matching**: We also introduce the BERT score [38] to evaluate the semantic similarity of generated scene descriptions with the annotated (ground-truth) ones. Precision measures the portion of semantically matched texts relative to the ground-truths, while recall is relative to the generated scene description. F1 is the combined measure of both the precision and recall. The higher values in precision, recall, and F1 score indicate better matching.
- **Quality**: This evaluation aims to gain a quality understanding regarding the results of (i) the VXI location descriptions, (ii) the VXI scene understanding and reasoning text, and (iii) the entire response (i.e., all the texts in the json format). We leverage GPT-4o-mini [5] to take in each of the above generated texts (estimations) and provide a score by comparing its quality against our given texts (ground-truths). In our implementation, we have incorporated the roles and tasks of GPT-4o-mini, the expected formats (e.g., strings), and the scoring requirements (e.g., correctness, informativeness) within our prompts to generate the scores (0–100 points). We first find the average quality scores of all the test VXI scenes, and the higher scores indicate the higher quality of the generated scene descriptions, understanding and reasoning texts, and overall responses. We then calculate the average score differences of the baseline and SOTA approaches against VXI-SUR to demonstrate the improvements.

• **Default Experimental Parameters and Environmental Settings.** In terms of scene augmentation, we empirically set $k = 3$ in order to balance between computational efficiency, and scene understanding and reasoning performance. Input images of the VXI scenes are of dimension $448 \times 448 \times 3$. We set the maximum dynamic patch of visual encoder in the InternVL backbone as 6. In our visual encoder of the MLLM, we set $m = 256$, $d = 896$, and the value of $n + m$ is upper bounded by 4,096 in our studies, and other settings follow the default ones in InternVL.

In our experimentation, we use a GPU server with AMD Ryzen Threadripper 3960X 24-Core CPU, 2×RTX8000 48GB GDDR5, and 128GB RAM for our scene-knowledge co-learning and fine-tuning. Another GPU server with AMD Ryzen Threadripper 3960X 24-Core CPU, 4×RTX3090 24GB GDDR5, and 128GB RAM is used for model inference (VXI scene understanding and reasoning). On both servers, we disable the flash attention [39], [40] and adopt FP16 precision for the model weights. We set the learning rate as 4e-5 with the weight decay rate of 0.01. We use 70% of all VXI scenes for model fine-tuning (with a total of 5 epochs and a batch size of 4), and the rest of VXI scenes for model inference.

B. Experimental Results

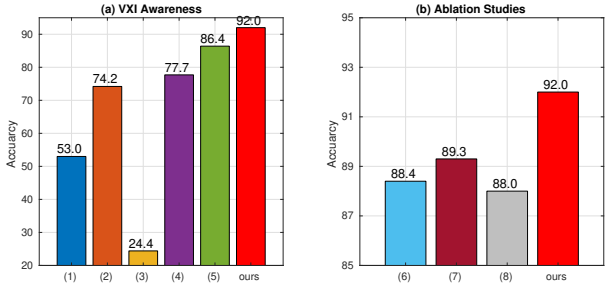

Figure 6: (a) VXI awareness regarding the OoI types; and (b) ablation studies regarding designs within VXI-SUR.

• **VXI Awareness.** We demonstrate the risk awareness of VXI-SUR in Fig. 6(a) in terms of the accuracy of VXI types. We can observe the low performance of (1) ChatGPT-4o-Mini, (2) Gemini-Flash, and (3) Claude-Haiku in terms of classifying the VXI types, implying the limited generalization ability of these models in VXI scene understanding and reasoning. Among these three models, (2) Gemini-Flash performs the best in terms of recognizing the VXI types. We can also observe that (4) InternVL, with effective knowledge distillation, shows better performance than (5) DeepSeek-VL in recognizing the VXI types (86.4% vs. 77.7%). Thanks to the scene-knowledge augmentation and co-learning designs, our VXI-SUR achieves overall better accuracy (on average 31.4% improvements) in classifying the VXI types, demonstrating its effectiveness in VXI awareness.

We show the model ablation study in Fig. 6(b) regarding comparison of VXI-SUR with respect to (6) without scene augmentation; (7) without knowledge augmentation; and (8) without VXI memory (i.e., without both scene and knowledge augmentation). We can observe that the (6) VXI-SUR w/o scene augmentation achieves even worse than (7) VXI-SUR w/o knowledge augmentation, implying the importance of peer VXI scenes for the VXI awareness. In the interest of space, in what follows, we focus on comparing our VXI-SUR with other baseline approaches (1)–(5).

• **Preciseness.** We further evaluate the preciseness of the generated scene description against the annotations (ground-truths) in Fig. 7. We can observe that (1) ChatGPT-4o-Mini, (2) Gemini-Flash, and (3) Claude-Haiku achieve overall low preciseness, since they have not fully captured and characterized the scene descriptions. Thanks to the provided designs of scene-knowledge augmentation and co-learning,

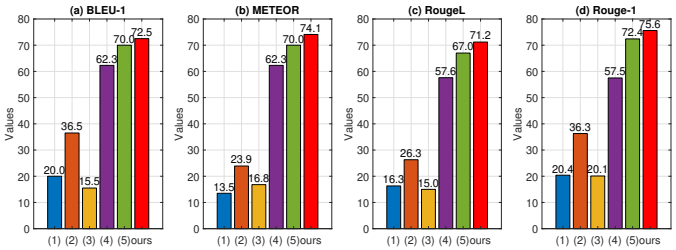

Figure 7: Preciseness evaluations of the scene description in terms of (a) BLEU-1; (b) METEOR; (c) RougeL; and (d) Rouge-1.

we can observe that VXI-SUR achieves on average 43.6%, 49.7%, 48.8%, and 45.3% improvements in terms of BLEU-1, METEOR, Rogue-L, and Rouge-1 compared with all the baseline and state-of-the-art approaches. In particular, we can see improvements of 8.7%, 10.7%, 12.5%, and 14.1% in the four metrics, on average, compared with DeepSeek-VL and InternVL.

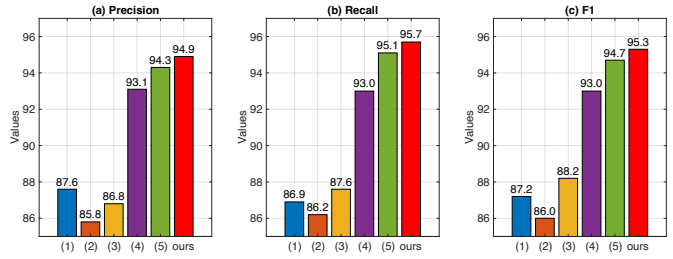

Figure 8: Semantic matching in terms of the BERT score for scene description: (a) precision; (b) recall; and (c) F1.

• **Semantic Matching.** We evaluate the BERT score based on the scene description and ground-truths in Fig. 8. We note that Claude-Haiku (3) outperforms Gemini-Flash (2) in terms of semantic matching, while the latter works better in terms of preciseness. Such discrepancy arises because Gemini-Flash tends to generate shorter sentences, which may favor BLEU-1 scores but not in semantic matching. Overall, our VXI-SUR demonstrates consistent improvements compared with other baselines in terms of scene descriptions.

• **Quality.** Figs. 9(a) and (b) illustrate the average differences of quality scores compared with VXI-SUR for (a) VXI location description as well as (b) VXI scene understanding and reasoning text. The higher differences indicate the more performance drops compared with our VXI-SUR. On average, we can observe that VXI-SUR achieves 20.6 and 16.5 points higher compared with all the baseline approaches in (a) and (b), respectively. We also show the average differences of quality scores compared with VXI-SUR for (c) the VXI responses. We can observe that all other models cannot meet the specific needs for VXI scene understanding and reasoning tasks. Augmentation of the peer scenes and knowledge helps VXI-SUR outperform all other approaches, on average by 18.8 points, in terms of VXI responses.

• **Case Studies.** We further illustrate an example of vehicle-to-vehicle interaction in Fig. 10, where the annotated ground-truth presents the vehicle-following decisions in response to a truck ahead. We compare the performance of VXI-SUR and DeepSeek-VL in terms of the generated VXI responses (blue) and the quality scores (yellow; with the explanations by GPT-4o-mini). We can observe that DeepSeek-VL misinterprets the VXI scenes and provides the incorrect decision types and scene understanding and reasoning texts. VXI-SUR, thanks to the VXI memory and augmentation, accurately understands and reasons about the VXI scenes, and hence yields high scores in our quality evaluations.

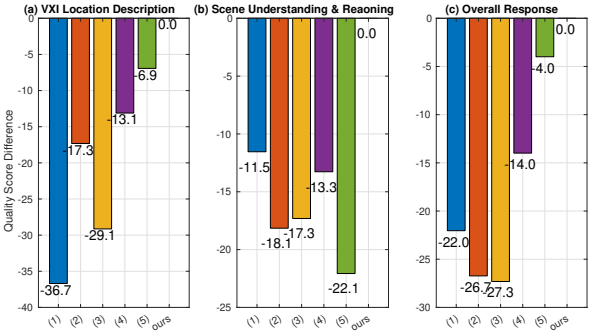

Figure 9: Evaluations of quality score differences compared with VXI-SUR regarding: (a) VXI location description; (b) VXI scene understanding and reasoning text; and (c) VXI response.

Vehicle-to-Vehicle Interaction Scene

Annotated Groundtruth

"**Scene Description**: "There is a black truck driving in the ego lane, away from the ego car",
"**VXI" Type**": "Vehicle",
"**VXI Location Description**": "ego lane",
"**Decision Type**": "Follow the vehicle ahead",
"**Scene Understanding and Reasoning Text**": "It\'s important to maintain a safe distance from the vehicle ahead of us, especially since there\'s a black truck driving in the ego lane away from our car. By following the vehicle ahead, we can keep a steady pace and avoid sudden braking or swerving that could lead to a collision. Make sure to monitor the speed and remain alert for any changes in traffic conditions. If the truck creates any hazards, be prepared to adjust your speed accordingly and keep a safe distance. Always prioritize safety by following traffic laws and being aware of your surroundings."

VXI-SUR

"Scene Description": "There is a black truck driving in the ego lane, away from the ego car.",
"VXI Type": "Vehicle",
"VXI Location Descrption": "ego lane",
"Decision Type": "Follow the vehicle ahead",
"Scene Understanding and Reasoning Text: "It\'s important to maintain a safe distance from the black truck in the ego lane, especially since it\'s moving away from us. Following the vehicle ahead will help us stay in our lane and avoid any potential hazards. Make sure to keep a safe following distance to ensure we have enough time to react if the truck changes speed or direction. Additionally, be aware of any vehicles around us, and signal your intentions if you need to change lanes or adjust your speed. Always prioritize safety and adhere to traffic laws."

'Score': 95,

'Reason': 'The predictions are nearly identical to the groundtruth in content, with minor wording differences. It is informative and follows the JSON format well, providing safety advice relevant to the situation.'

DeepSeek-VL

"Scene Description": "There is a black truck stopped in front of the ego car in the ego lane, because of a red traffic light ahead.",
"VXI Type": "Vehicle",
"VXI Location Description": "ego lane",
"Decision Type": "(must) Stop",
"Scene Understanding and Reasoning Text": "Since there is a black truck stopped directly in front of us due to a red traffic light, it is essential that we stop ahead of that vehicle to adhere to traffic laws. Continuing to drive without stopping could lead to a collision, as that truck may not be fully in motion or be preparing to move. Make sure to apply the brakes smoothly and remain alert, maintaining a safe distance while waiting for the light to change. Always adhere to traffic signals to ensure public safety on the road."

'Score': 20,
'Reason': 'The prediction is incorrect as it describes a stopped truck, while the ground truth describes a driving truck. The suggestion and instruction differ significantly from the ground truth, which impacts the informativeness for the driver. The JSON format is followed.'

Figure 10: Illustration of a vehicle-to-vehicle interaction scene and comparison of VXI-SUR with DeepSeek-VL.

VII. Conclusion

We have designed VXI-SUR to overcome the challenges in gaining semantic and structured scene understanding and reasoning about the complex vehicle-to-X interaction (VXI) scenes. We have developed the VXI memory with scene-knowledge augmentation mechanism which augments the VXI

scenes (reflecting the peer scenes that are similar to the target VXI scene) and knowledge (regarding the objects of interest as well as their relative positions and closeness). Such VXI memory strengthens structured MLLM fine-tuning. We have designed a scene-knowledge co-learning mechanism, and performed comprehensive evaluations of the resulting system based on a real-world open-source dataset named DRAMA. We have conducted extensive experimental studies upon VXI-SUR, and validated its effectiveness and usefulness in terms of the VXI awareness, description preciseness, semantic matching, and quality in understanding and reasoning VXI scenes.

ACKNOWLEDGMENT

This project is supported, in part, by the National Science Foundation (NSF) under Grant No. 2239897, Google Research Scholar Program Award, and NVIDIA Applied Research Accelerator Program Award. We would like to thank Honda Research Institute (HRI), US, and Connecticut Transportation Institute (CTI) for their support of our research projects.

REFERENCES

[1] Y. Zhang, Q. Ma, J. Qu, and R. Zhou, "Effects of Driving Style on Takeover Performance During Automated Driving: Under the Influence of Warning System Factors," *Applied Ergonomics*, vol. 117, p. 104229, 2024.

[2] M. Tabatabaie and S. He, "Driver maneuver interaction identification with anomaly-aware federated learning on heterogeneous feature representations," *Proc. ACM IMWUT*, vol. 7, no. 4, Jan. 2024.

[3] M. Tabatabaie, S. He, H. Wang, and K. G. Shin, "Beyond "Taming Electric Scooters": Disentangling Understandings of Micromobility Naturalistic Riding," *Proc. ACM IMWUT*, vol. 8, no. 3, Sep. 2024.

[4] M. Tabatabaie, S. He, and K. G. Shin, "Cross-Modality Graph-Based Language and Sensor Data Co-Learning of Human-Mobility Interaction," *Proc. ACM IMWUT*, vol. 7, no. 3, pp. 1–25, 2023.

[5] OpenAI, "GPT-4o-mini," gpt-4o-mini-2024-07-18, 2024, large Language Model.

[6] Google, "Gemini-Flash," gemini-1.5-flash, 2024, large Language Model.

[7] Anthropic, "Claude-Haiku," claude-3-haiku-20240307, 2024, large Language Model.

[8] H. Lu, W. Liu, B. Zhang, B. Wang, K. Dong, B. Liu, J. Sun, T. Ren, Z. Li, Y. Sun *et al.*, "Deepseek-VL: Towards Real-World Vision-Language Understanding," *arXiv preprint arXiv:2403.05525*, 2024.

[9] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu *et al.*, "InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks," in *Proc. IEEE/CVF CVPR*, 2024, pp. 24 185–24 198.

[10] M. J. Prohn and B. Herbig, "Potentially Critical Driving Situations During "Blue-light" Driving: A Video Analysis," *Western Journal of Emergency Medicine*, vol. 24, no. 2, p. 348, 2023.

[11] L. Wen, D. Fu, X. Li, X. Cai, T. Ma, P. Cai, M. Dou, B. Shi, L. He, and Y. Qiao, "Dilu: A Knowledge-driven Approach to Autonomous Driving with Large Language Models," *arXiv preprint arXiv:2309.16292*, 2023.

[12] S. Malla, C. Choi, I. Dwivedi, J. H. Choi, and J. Li, "DRAMA: Joint Risk Localization and Captioning in Driving," in *Proc. IEEE/CVF CVPR*, 2023, pp. 1043–1052.

[13] M. Xu, W. Yin, D. Cai, R. Yi, D. Xu, Q. Wang, B. Wu, Y. Zhao, C. Yang, S. Wang *et al.*, "A survey of resource-efficient LLM and multimodal foundation models," *arXiv preprint arXiv:2401.08092*, 2024.

[14] M. Reid, N. Savinov, D. Teplyashin, D. Lepikhin, T. Lillicrap, J.-b. Alayrac, R. Soricut, A. Lazaridou, O. Firat, J. Schrittwieser *et al.*, "Gemini 1.5: Unlocking Multimodal Understanding Across Millions of Tokens of Context," *arXiv preprint arXiv:2403.05530*, 2024.

[15] L. Caruccio, S. Cirillo, G. Polese, G. Solimando, S. Sundaramurthy, and G. Tortora, "Claude 2.0 Large Language Model: Tackling a real-world classification Problem with a New Iterative Prompt Engineering Approach," *Intelligent Systems with Applications*, vol. 21, p. 200336, 2024.

[16] S. Hu, Y. Tu, X. Han, C. He, G. Cui, X. Long, Z. Zheng, Y. Fang, Y. Huang, W. Zhao *et al.*, "MiniCPM: Unveiling the Potential of Small Language Models with Scalable Training Strategies," *arXiv preprint arXiv:2404.06395*, 2024.

[17] R. Pi, L. Yao, J. Gao, J. Zhang, and T. Zhang, "PerceptionGPT: Effectively fusing visual perception into LLM," in *Proc. IEEE/CVF CVPR*, 2024, pp. 27 124–27 133.

[18] T. Choudhary, V. Dewangan, S. Chandhok, S. Priyadarshan, A. Jain, A. K. Singh, S. Srivastava, K. M. Jatavallabhula, and K. M. Krishna, "Talk2BEV: Language-enhanced Bird's-eye View Maps for Autonomous Driving," *arXiv preprint arXiv:2310.02251*, 2023.

[19] X. Ding, J. Han, H. Xu, X. Liang, W. Zhang, and X. Li, "Holistic Autonomous Driving Understanding by Bird's-Eye-View Injected Multi-Modal Large Models," in *Proc. IEEE/CVF CVPR*, 2024, pp. 13 668–13 677.

[20] S. Fang, J. Liu, M. Ding, Y. Cui, C. Lv, P. Hang, and J. Sun, "Towards interactive and learnable cooperative driving automation: a large language model-driven decision-making framework," *IEEE Transactions on Vehicular Technology*, pp. 1–12, 2025.

[21] P. Taveekitworachai, P. Suntichaikul, C. Nukoolkit, and R. Thawonmas, "Speed Up! Cost-Effective Large Language Model for ADAS Via Knowledge Distillation," in *Proc. IEEE IV*, 2024, pp. 1933–1938.

[22] Y. Huang, J. Sansom, Z. Ma, F. Gervits, and J. Chai, "DriVLMe: Enhancing LLM-based Autonomous Driving Agents with Embodied and Social Experiences," *arXiv preprint arXiv:2406.03008*, 2024.

[23] H. Sha, Y. Mu, Y. Jiang, L. Chen, C. Xu, P. Luo, S. E. Li, M. Tomizuka, W. Zhan, and M. Ding, "LanguageMPC: Large language models as Decision Makers for Autonomous Driving," *arXiv preprint arXiv:2310.03026*, 2023.

[24] Honda, "Honda Research Institute DRAMA Dataset," https://usa.honda-ri.com/drama, 2024.

[25] OpenAI, "GPT-3.5," gpt-3.5-turbo-0125, 2021, large Language Model.

[26] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large Language Models are Zero-shot Reasoners," *Proc. NeurIPS*, vol. 35, pp. 22 199–22 213, 2022.

[27] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning Transferable Visual Models from Natural Language Supervision," in *Proc. ICML*. PMLR, 2021, pp. 8748–8763.

[28] J. J. Pan, J. Wang, and G. Li, "Survey of Vector Database Management Systems," *The VLDB Journal*, vol. 33, no. 5, pp. 1591–1615, 2024.

[29] J. Wang, X. Yi, R. Guo, H. Jin, P. Xu, S. Li, X. Wang, X. Guo, C. Li, X. Xu *et al.*, "Milvus: A Purpose-Built Vector Data Management System," in *Proc. ICMD*, 2021, pp. 2614–2627.

[30] D. Reis, J. Kupec, J. Hong, and A. Daoudi, "Real-time Flying Object Detection with YOLOv8," *arXiv preprint arXiv:2305.09972*, 2023.

[31] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth Anything V2," *arXiv:2406.09414*, 2024.

[32] X. Ding, J. Han, H. Xu, W. Zhang, and X. Li, "HiLM-D: Towards High-Resolution Understanding in Multimodal Large Language Models for Autonomous Driving," *arXiv preprint arXiv:2309.05186*, 2023.

[33] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank Adaptation of Large Language Models," *arXiv preprint arXiv:2106.09685*, 2021.

[34] Y. Zhao, J. Huang, J. Hu, X. Wang, Y. Mao, D. Zhang, Z. Jiang, Z. Wu, B. Ai, A. Wang, W. Zhou, and Y. Chen, "SWIFT:A Scalable LightWeight Infrastructure for Fine-Tuning," 2024.

[35] K. Papinesi, "Bleu: A method for Automatic Evaluation of Machine Translation," in *Proc. ACL*, 2002, pp. 311–318.

[36] S. Banerjee and A. Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," in *Proc. ACL Workshop*, 2005, pp. 65–72.

[37] L. Chin-Yew, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Text Summarization Branches Out*, 2004.

[38] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," in *Proc. ICLR*, 2020.

[39] T. Dao, D. Y. Fu, S. Ermon, A. Rudra, and C. Ré, "FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness," in *Proc. NeurIPS*, 2022.

[40] T. Dao, "FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning," in *Proc. ICLR*, 2024.