Visual Grounding via Heterogeneous Representation Learning and Hierarchical Reasoning of Human-to-Vehicle Commands

Hao Wang[1] Suining He[1]* Kang G. Shin[2]
[1]Ubiquitous & Urban Computing (U^2C) Lab, University of Connecticut
[2]Real-Time Computing Lab (RTCL), University of Michigan–Ann Arbor

***Abstract*— With the proliferation of autonomous vehicles (AVs) and their increasing interaction and communication with the riders, how to ground or locate the visual objects of interests (OoIs), such as the concerned pedestrians and other traffic participants, based on the human riders' natural language and communication (e.g., vocal commands), is essential for increasing the efficiency, effectiveness, and reliability/safety of AVs in following the riders' reasonable commands and preferences. There are several technical challenges to achieve visual grounding for such human-to-vehicle commanding (HVC) scenes, including (1) how to fuse heterogeneous sensor modalities — i.e., visual object information, textual contexts, and situation awareness (say, obtained from the light detection and ranging); (2) how to discern the opaque commands in the human natural language; and (3) how to reason about the relative positions of the OoIs within the visual modality.**

To meet these challenges, we propose VIGOR, a VIsual Grounding approach based on heterogeneous mOdality learning and hierarchical Reasoning for HVC scenes. First, we design a heterogeneous modality learning approach in order to incorporate the visual, textual, and situational modalities, and learn their cross-modality representations to identify important information for visual grounding. Then, VIGOR performs hierarchical reasoning of objects and context levels, and differentiates the OoIs in the complex traffic environments that relate to the natural language commands. Finally, we conduct extensive experimental studies on a total of 12,037 HVC scenes, demonstrating VIGOR to achieve higher accuracy than the state-of-the-art approaches (by 14.81% on average) in terms of the Intersection over Union (IoU) in grounding the OoIs in the complex (including low-visibility) HVC scenes.

I. INTRODUCTION

This paper focuses on grounding (locating) the objects of interest (OoIs) that are interacting with the focal autonomous vehicle (AV) and need specific actions or responses, which has become essential for various human-to-vehicle commanding (HVC), and collaboration applications [1], [2]. For instance, the rider(s) in the focal AV (e.g., a robotaxi of Waymo or Apollo Go) can provide specific and reasonable natural language command or guidance on a certain street or neighborhood when a route change or added stop is needed, during parallel or curbside parking, as well as for raised caution of specific traffic participants (e.g., pedestrians).

Our HVC studies in this paper focus on the visual modality — i.e., the image frames from the cameras available in the perception modules of AVs — since the human riders likewise discern and locate the OoIs through their bare eyes. Such OoIs often include other traffic participants (e.g., other vehicles, cyclists), pedestrians, and infrastructures (e.g., traffic lights, crosswalks). Their accurate grounding given HVC can enhance the perception of AVs (and benefit the subsequent prediction [3], planning, and control tasks [4]), and improve collaborativeness of AVs with human riders.

We particularly note that the HVC scenes differ from conventional visual grounding scenarios. In the HVC scenes, the free-form commands tend to be short, and highly context-dependent in terms of object referral (e.g., "park behind the blue car on the right"), in contrast to the grounding [5], [6]. Thus, we need to develop an accurate visual grounding system for HVC scenes, and discern and locate the relevant OoIs based on the complex free-form natural language by the human riders. Such a system can be integrated with the advanced driver assistance system (ADAS) or the human-vehicle interfaces to support the human-vehicle interaction and collaboration.

From prior arts and practices, we find it necessary to address the following two challenges toward a practical and ubiquitous visual grounding system for HVC.

(1) How to realize the cross-modality sensor fusion for visual grounding in complex traffic environments? Unlike prior studies, our HVC scene needs to align the textual modalities with the visual ones toward semantic correspondence of visual OoIs and textual commands. Given the complex (say, urban) traffic environments, existing visual grounding approaches (such as those in [5], [6], [7]) may not necessarily discern and locate all the OoIs, especially under the low-light (e.g., lack of streetlights at night) or low-visibility conditions (say, heavy rain or bright light). A viable workaround is through a multi-modality setting, such as fusing visual and textual modalities with the additional situational information [8]. These modalities can be images, human natural language, and point clouds from the Light Detection and Ranging (LiDAR). However, realizing a cross-modality formulation is non-trivial as the challenge comes from the heterogeneity (e.g., quality, granularity) and discrepancy (e.g., scopes, views) of modalities. For instance, the brief human-to-vehicle commands may differ from the information-rich visual images. In the meantime, the situational modality (e.g., the LiDAR point clouds) may not necessarily align with the AV riders' visual perception (say, the dash view), as well as their intentions in grounding specific OoIs.

(2) How to mitigate the impacts of AV riders' opaque commands in discerning the referred OoIs and their relative positions? In the HVC scenes in complicated traffic environments, AV riders may not always provide comprehensive or specific references regarding the crucial

*Contact Author: Suining He (suining.he@uconn.edu)

OoIs. These may be due to various environmental and human factors such as limited response time, performing non-driving related tasks, etc. Let us take the human-assisted AV parking as an instance. A natural language command, such as "this is the place and pull up next to that adult over there", may not necessarily provide sufficient constraints regarding the specific individual within the visual modality, particularly in a complex traffic environment with multiple relevant objects. Moreover, grounding the OoIs involved with a certain mobility status (e.g., a stalled car) may often need specific *contexts* from the cross-modality fusion — across visual, textual, and situational information — to capture the physical dynamics of the OoIs for performing subsequent tasks. Mapping the exact OoIs from the natural language commands to the features returned from the heterogeneous modalities, and capturing the relative closeness of the OoIs (i.e., reasoning rather than guessing), is also essential for fine-grained and correct visual grounding.

To overcome the above challenges, we have designed **VIGOR**, a **VI**sual **G**rounding approach based on heterogeneous m**O**dality learning and hierarchical **R**easoning for HVC scenes. Specifically, we make the following technical contributions.

1) We have designed a heterogeneous representation learning approach. VIGOR incorporates and captures the complex representations from the textual (e.g., commands in human natural language), visual (e.g., images from the AV's front-view camera), and situational modality (e.g., LiDAR) information for subsequent accurate and fine-grained grounding of OoIs. We have designed visual-textual attention (`VTA`) and situational attention with adaptive sampling (`SAAS`) to fuse and extract the cross-modality feature information that is useful for discerning the OoIs, and reduce the unnecessary one for grounding (e.g., redundant road environment information from LiDAR).
2) To further differentiate the OoIs for accurate visual grounding and effective HVC, we have designed a hierarchical reasoning mechanism for the grounding tasks by VIGOR. Specifically, our visual grounding integrates the object- and context-level reasoning from the complex HVC scenes. We have designed the part-of-speech (POS) tagging to strengthen hierarchical reasoning of objects and the contexts of the scenes. For instance, given a command of "follow the car in front of us in the left lane", the object-level reasoning will relate key nouns and actions like "car", "lane", and "follow", while the context-level reasoning will involve important contexts (e.g., adjectives, adverbs) like "front" and "left". Such a hierarchical reasoning design yields more accurate and fine-grained grounding results than conventional grounding approaches (such as those in [6], [9]).
3) We have conducted extensive data-driven experimental studies on a total of 12,037 HVC scenes or samples (with the size of 23.83Gb in total) adapted from Talk2Car and NuScenes datasets [10], [11]. We have studied various HVC scenes including those in the nighttime and rainy conditions. VIGOR is shown to make more than 15% accuracy improvements (in terms of Intersection over Union or IoU) over other baselines and state-of-the-art (SOTA) approaches [12], [9] (see `github.com/middleflames/VIGOR` for more details).

II. Related Work

Human-to-Vehicle Commanding. Effective communication in the human natural language during the HVC is essential for enabling various human-in-the-loop decision support tasks [13], [14], [7], such as autonomous parking or language-guided navigation [15]. Furthermore, HVC is essential for improving the receptivity and trustworthiness of AVs by a broader spectrum of human riders, including those visually impaired. In this paper, we focus on grounding the OoIs the focal vehicle is interacting with, which can serve as a case study to capture the important cross-modality learning and hierarchical reasoning insights for HVC scenes.

Visual Grounding and Object Referral. By locating the most relevant objects in the visual modality (e.g., image) based on the textual modality (e.g., natural language), visual grounding (or object referral) aims to understand both modalities, and align their correspondences. For instance, `AttnGrounder` [12] accounted for different resolutions of images with the attention mechanism to align the visual and textual information for grounding the OoIs. `VL-BERT` [5] provided a vision–language architecture based on Transformer [16]. `CMSVG` [7] accounted for the representations from textual and visual modalities, and leveraged the modality similarity for visual grounding. General referring expression comprehension [17] also adopted Transformer [16] for fusing visual and textual modalities [6]. `X-VLM` [18] explored the relationship between text and visual concepts in images. `HiVG` [19] studied hierarchical low-rank adaptation for concise and efficient visual grounding. `VGNet` [20] proposed layer-wise vision-text interaction to progressively highlight the target objects. `SimVG` [21] and C^3VG [22] respectively decouple visual-linguistic features and impose coarse-to-fine consistency constraints in their frameworks. These visual grounding approaches designed for general vision task, however, may not adapt well to the complex traffic environments where the OoIs can be highly interactive and heterogeneous. The resulting HVC may not provide comprehensive referrals on critical objects, and hence degrade these approaches' effectiveness in grounding the objects.

VIGOR advances in the following aspects. (a) It provides hierarchical levels of reasoning regarding the objects and contexts that are concerned with the HVC scenes, beyond the conventional grounding approaches [23], and thus yields fine-grained grounding results in complex traffic environments. (b) Given the heterogeneous (i.e., visual, textual, and situational) modalities, we have designed and validated our heterogeneous representation learning task upon the cross-

modality features that are essential for comprehending the HVC scenes and the engaged OoIs.

III. Overview of VIGOR

A. Problem Definition

Given the heterogeneous modalities in the HVC scenes, i.e., $\mathbf{V}_t$, $\mathbf{T}_t$, and $\mathbf{L}_t$, at timestamp t, we want to find a function $f_{\boldsymbol{\theta}}(\mathbf{V}_t, \mathbf{T}_t, \mathbf{L}_t)$, with all the model parameters denoted as $\boldsymbol{\theta}$, that returns the bounding box $\widehat{\mathbf{y}}_t$ of the OoIs that has the smallest differences from the ground-truth $\mathbf{y}_t$, as well as the object-level and context-level POS tags for hierarchical reasoning.

B. Framework Overview

Fig. 1 illustrates the system framework of VIGOR which consists of three major tasks toward visual grounding during the HVC: (1) Data Pre-processing and Modality Preparation, (2) Heterogeneous Representation Learning, and (3) Hierarchical Reasoning and Visual Grounding.

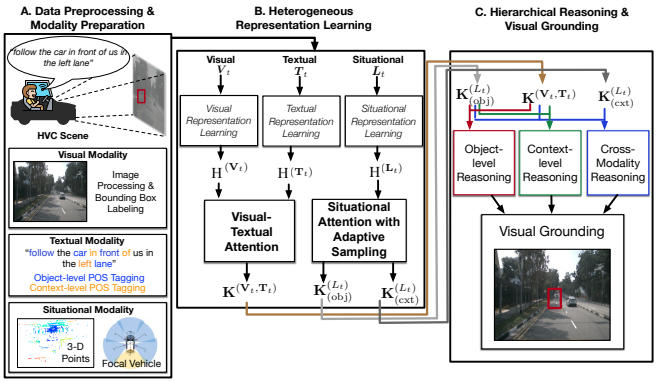

Fig. 1: Overview of the system framework of VIGOR.

1) Data Pre-processing and Modality Preparation: We focus on the three heterogeneous modalities: textual, visual, and situational modalities. Specifically, the textual modality refers to the human natural language commands like "park behind that grey van on the other side of the road". The visual modality accounts for the scenes captured by, for example, the front-view camera of the focal vehicle. The situational modality refers to the key LiDAR frame relevant to the HVC scene.

For our prototype development, we have structured and adapted HVC scenes (samples) from the Talk2Car [11] dataset. The Talk2Car is built on the NuScenes [10] data, and provides the natural language command to the focal vehicle toward the destination of interest when interacting with various traffic participants. We note that the original Talk2Car dataset only contains the visual (images) and textual (commands) modalities, and we retrieve the original LiDAR frames from NuScenes as the situational modality that correspond to the HVC scenes. A total of 12,037 HVC scenes are prepared for our VIGOR studies. In what follows, we present the details of data processing for each of the modalities involved.

(a) *Visual Modality.* For our visual grounding, we have performed contrast maximization, histogram equalization, random brightness, and sharpness changes upon each image $\mathbf{V}_t$ at timestamp t. We form the processed image frame at timestamp t into a tensor $\mathbf{V}_t \in \mathbb{R}^{W \times H \times C}$, where W, H, and C, respectively, represent the width, height, and number of channels of each image. We also label the bounding boxes (normalized with respect to the height and width of the image) of the important OoIs that concern the AV rider's natural language commands during HVC. In particular, we have identified a total of 8,089 HVC scenes that concern other vehicles, 2,990 scenes that concern pedestrians, and 940 that concern infrastructures. Note that one HVC scene may involve more than one type of OoIs.

(b) *Situational Modality.* Toward inclusion of situation awareness regarding the OoIs near the focal vehicle, we have further taken in the LiDAR for the cross-modality learning in VIGOR. Specifically, for each visual modality measured, we have taken into account the LiDAR frames (consisting of the 3-D point cloud) whose timestamps are the closest to the timestamp of the image frame. We take the average of the 10 closest LiDAR frames to form each LiDAR frame at the timestamp t for VIGOR training (the sampling frequency of LiDAR from the nuScenes dataset is 20Hz [10]). We denote the resulting key LiDAR frame, which is represented as a voxel (i.e., the regular 3-D grid; counterpart of a pixel in the 2-D space) for the subsequent heterogeneous representation learning at timestamp t, as $\mathbf{L}_t \in \mathbb{R}^{N_{\text{(voxel)}} \times 3}$, where $N_{\text{(voxel)}}$ represents the number of voxels in the point cloud, and each element corresponds to a 3-D coordinate (relative to the local coordinate system of the focal vehicle).

(c) *Textual Modality.* For the natural language commands during each HVC scene, we first perform the tokenization to transform the words into tokens (units). Then, we perform the Part of Speech (POS) tagging upon the tokens to label the groups of words that relate to either the objects of interests (e.g., other traffic participants) or the contexts of the scenes (such as "drop me off there"). Specifically, we prepare the following two levels of tagging in textual modalities for hierarchical understanding of the HVC scenes and performing visual grounding.

- *Object-level POS Tagging:* Within each natural language command, we tag the nouns that correspond to the concerned OoIs (e.g., cars, pedestrians), and verbs that correspond to the relevant actions (e.g., "drop off") in this HVC scene.
- *Context-level POS Tagging:* The human natural language commands, such as "drop me off there", may not necessarily provide all the nouns or verbs relevant to the HVC scene. Therefore, we further take into account the contexts relevant to the scene, such as the adverbs (e.g., related to the manner and degree), adjectives (e.g., related to the size or color of the OoIs), and adpositions (e.g., related to the spatial and temporal relations). For instance, despite the absence of nouns in the AV rider's command "drop me off there", we can further tag the semantics and relative positions to strengthen the model's comprehension upon the contexts.

With all of the above, VIGOR performs the hierarchical reasoning in terms of object and context levels. In order to contrast the POS tags that are of interests by VIGOR in each level, we leverage the label of "other" to denote all other unrelated tokens. For instance, for the object level, we label the tokens except for those that correspond to the nouns and verbs (say, labeled as 1 and 2) with "other" (with 0). Similarly, when performing the context-level reasoning, all tokens except for those corresponding to the adjectives (Adj), adpositions (Adp), and adverbs (Adv), which are labeled as 1, 2, and 3, are assigned as "other" (with 0).

We illustrate two examples of HVC scenes in Fig. 2 as well as their corresponding modalities (visual, situational, and textual). We show one HVC scene (a) where the driver communicates with the command of "slow down and turn in here and park near that man on the sidewalk", and one low-visibility scene (b) (i.e., rainy conditions) with the command of "follow the car in front of us in the left lane". Currently, we have identified a total of 2,990 scenes with pedestrians, 8,089 with vehicles, and 940 with others such as infrastructures (say, traffic signs and sidewalks). In terms of textual modalities, we have labeled a total of 50,741 object-level POS tags and 28,792 context-level POS tags.

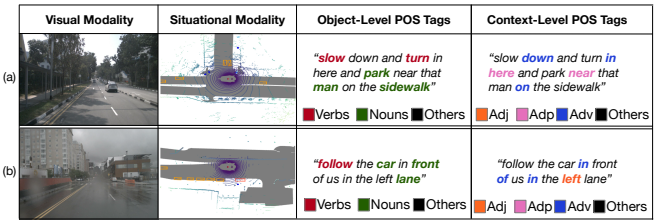

Fig. 2: Two HVC examples and their modalities.

2) Heterogeneous Representation Learning: Given the processed heterogeneous modality data, we have further incorporated multiple encoder models to encode their representations and generate the feature embeddings, respectively. The key idea is to structure the heterogeneous modalities into the forms that can capture the OoIs, contexts, and AV riders' intentions, and further feed them for hierarchical reasoning and visual grounding. We will perform operations of visual-textual attention, and situation attention with adaptive sampling. Note that despite the current prototype based on the above-mentioned models, our framework is general enough to be integrated within other encoder ones.

3) Hierarchical Reasoning and Visual Grounding: Given the generated feature embeddings from the models regarding the heterogeneous modalities, VIGOR further learns the correlations across the representations of the textual and visual modalities. In addition, VIGOR takes in the situational modality representation, and further conducts object-level, context-level, and cross-modality reasoning upon the textual modality (POS tags).

IV. Heterogeneous Representation Learning

The key idea of our heterogeneous representation learning task is to extract the respective representations from the input modalities (visual, textual, and situational) that can be further fused for the challenge (1) in Section I. Through the extraction, VIGOR can focus on establishing the attention mechanism and capturing the cross-modality interactions among the modalities.

• **Visual, Textual, & Situational Representation Learning.** In what follows, we present the details of each encoder model for the visual, textual, and situational modalities.

For the visual representation learning, we adapt the Swin Transformer [24] in this prototype study. Specifically, VIGOR first splits each image (frame) $\mathbf{V}_t$ at the timestamp t into a total of N_{patch} patches (i.e., containers of pixels extracted from the input image) [24]. We have adopted three consecutive sets of Swin Transformer blocks in order to generate the tokenized embeddings or representations, i.e., $\mathbf{H}^{(\mathbf{V}_t)} \in \mathbb{R}^{N_{\text{tok}}^{(\mathbf{V})} \times N_{\mathbf{V}}}$, where $\mathbb{R}^{N_{\text{tok}}^{(\mathbf{V})}}$ represents the number of tokens in visual modality, and $N_{\mathbf{V}}$ represents the dimension for each token.

To conduct the textual representation learning, given the human natural language command $\mathbf{T}_t$, we leverage `BERT` for tokenized embeddings, denoted as $\mathbf{H}^{(\mathbf{T}_t)} \in \mathbb{R}^{N_{\text{tok}}^{(\mathbf{T})} \times N_{\mathbf{T}}}$, where $N_{\text{tok}}^{(\mathbf{T})}$ and $N_{\mathbf{T}}$, respectively, represent the number of tokens in textual modality, and the number of dimension for each token (the dimensions of hidden states within `BERT`).

To perform the situational representation learning, we have incorporated `PillarNet` [25] to encode the 3-D point cloud, which transforms the input situational modality $\mathbf{L}_t$ (3-D points) into pillars and encodes them through convolution and de-convolution operations [25]. Then, we further project, through a linear projection layer, our situational representation embeddings into the tokenized representation, denoted as $\mathbf{H}^{(\mathbf{L}_t)} \in \mathbb{R}^{N_{\text{tok}}^{(\mathbf{L})} \times N_{\mathbf{L}}}$, where $N_{\text{tok}}^{(\mathbf{L})}$ and $N_{\mathbf{L}}$, respectively, represent the number of tokens in situational modality, and the dimension for each token (dimensions of hidden states from the final output layer of the `PillarNet` model).

In order to capture the modality correlations across the heterogeneous feature representations encoded from our encoder models, we have designed the visual-textual attention (`VTA`) as well as situational attention with adaptive sampling (`SAAS`).

• **Visual-Textual Attention (`VTA`).** The key idea of our visual-textual attention is to fuse the tokenized representations from the encoder models, Swin Transformer and `BERT`, based on the point-wise attention [16]. We note that the visual and textual modalities, when tokenized into embeddings, are temporally aligned and dense in representations.

Specifically, the output $\mathbf{K}^{(\mathbf{V}_t,\mathbf{T}_t)} \in \mathbb{R}^{N_{\text{tok}}^{(\mathbf{T})} \times N_{\mathbf{T}}}$ from visual-textual attention (denoted as `VTA`) from both the visual and textual modalities, i.e., $\mathbf{H}^{(\mathbf{V}_t)}$ and $\mathbf{H}^{(\mathbf{T}_t)}$, is given by

$$\mathbf{K}^{(\mathbf{V}_t,\mathbf{T}_t)} = \texttt{VTA}\left(\mathbf{H}^{(\mathbf{T}_t)}, \mathbf{H}^{(\mathbf{V}_t)}\right)$$
$$= \texttt{softmax}\left(\left(\mathbf{H}^{(\mathbf{T}_t)}\mathbf{W}_{\text{VTA}}^{(\text{query})}\right)\left(\mathbf{H}^{(\mathbf{V}_t)}\mathbf{W}_{\text{VTA}}^{(\text{key})}\right)^{\top}\right) \cdot \mathbf{H}^{(\mathbf{V}_t)}\mathbf{W}_{\text{VTA}}^{(\text{value})},$$

where $\mathbf{W}_{\text{VTA}}^{(\text{query})} \in \mathbb{R}^{N_{\mathbf{T}} \times N_{\mathbf{T}}}$, $\mathbf{W}_{\text{VTA}}^{(\text{key})} \in \mathbb{R}^{N_{\mathbf{V}} \times N_{\mathbf{T}}}$, and $\mathbf{W}_{\text{VTA}}^{(\text{value})} \in \mathbb{R}^{N_{\mathbf{V}} \times N_{\mathbf{T}}}$. In our `VTA`, we consider the embeddings from the textual modality as the query in the attention mechanism against the visual modality, since the human

natural language commands in HVC can provide the referral clues to the visual grounding results.

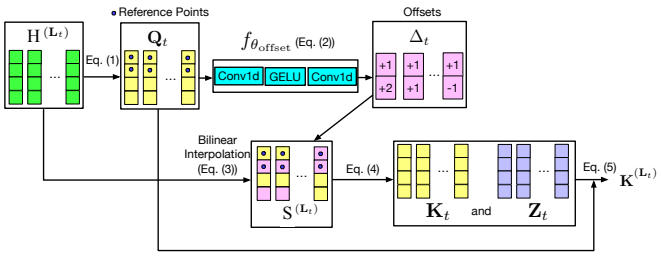

Fig. 3: Illustration of our SAAS mechanism.

• **Situational Attention with Adaptive Sampling (SAAS).** Note that the situational modality, particularly the LiDAR data studied in our prototype, provides panoramic views, and thus more feature information than necessary (e.g., irrelevant road environments, houses) for discerning the OoIs in the HVC scenes. Our visual-textual attention fuses the visual clues and textual commands that are largely related to the dash (forward) scope of the focal vehicle and the human rider. In order to adaptively align the scope of our situational modality with the other two modalities, we have designed situational attention with adaptive sampling, as illustrated in Fig. 3. The key idea of our situational attention with adaptive sampling (SAAS) is to focus on the subset of the input situational modality as the *reference points*, and resample the features from them or their nearby points, thus reducing the entire scope from the LiDAR point cloud.

Specifically, we let N_{ref} be the number of reference points, and first find the query $\mathbf{Q}_t \in \mathbb{R}^{N_{\text{tok}}^{(\mathbf{L})} \times N_{\text{ref}}}$ for VIGOR by projecting the tokenized embeddings $\mathbf{H}^{(\mathbf{L}_t)}$ from the situational modality representation learning, i.e.,

$$\mathbf{Q}_t = \mathbf{H}^{(\mathbf{L}_t)} \mathbf{W}^{(\text{query})}, \tag{1}$$

where $\mathbf{W}^{(\text{query})} \in \mathbb{R}^{N_{\mathbf{L}} \times N_{\text{ref}}}$ represents the weight matrix of the projection.

To let VIGOR focus on the important reference points that are actually relevant to OoIs within the HVC scenes, we calculate a set of offsets, denoted as $\mathbf{\Delta}_t \in \mathbb{R}^{N_{\text{ref}}}$, that consists of a total of N_{ref} scalars to help shift VIGOR's focuses (sampling points) upon the tokenized embeddings of the situational modality, i.e.,

$$\mathbf{\Delta}_t = f_{\boldsymbol{\theta}_{\text{offset}}}(\mathbf{Q}_t), \tag{2}$$

where $f_{\theta_{\text{offset}}}(\cdot)$ serves as an auxiliary function (consisting of two stacked 1-D convolutions with the GELU activation between them in our prototype) to learn and capture how far away each sampling point needs to be shifted from a reference point.

Let $\mathbf{E}_t \in \mathbb{R}^{N_{\text{ref}}}$ be the indices of the reference points in the situational modality. The resampled embeddings from the tokenized embeddings $\mathbf{H}^{(\mathbf{L}_t)} \in \mathbb{R}^{N_{\text{tok}}^{(\mathbf{L})} \times N_{\mathbf{L}}}$ is given by the bilinear interpolation [26], denoted as $\Phi(\cdot)$, with the resampled points from $\mathbf{H}^{(\mathbf{L}_t)}$ that are shifted based on indices $\mathbf{E}_t + \mathbf{\Delta}_t$, i.e.,

$$\mathbf{S}^{(\mathbf{L}_t)} = \Phi\left(\mathbf{H}^{(\mathbf{L}_t)}; \mathbf{E}_t + \Delta_t\right). \tag{3}$$

Given above, we let the key and value matrices, denoted as $\mathbf{K}_t \in \mathbb{R}^{N_{\mathbf{L}} \times N_{\text{ref}}}$ and $\mathbf{Z}_t \in \mathbb{R}^{N_{\mathbf{L}} \times N_{\text{ref}}}$, of our SAAS mechanism be

$$\mathbf{K}_t = \mathbf{S}^{(\mathbf{L}_t)} \mathbf{W}^{(\text{key})}, \quad \mathbf{Z}_t = \mathbf{S}^{(\mathbf{L}_t)} \mathbf{W}^{(\text{value})}, \tag{4}$$

where $\mathbf{W}^{(\text{key})}$, $\mathbf{W}^{(\text{value})} \in \mathbb{R}^{N_{\text{ref}} \times N_{\text{ref}}}$ represent the weight matrices for the key and value in the situational attention. Finally, our SAAS mechanism for the timestamp t returns the learned representation, denoted as $\mathbf{K}^{(\mathbf{L}_t)}$, i.e.,

$$\mathbf{K}^{(\mathbf{L}_t)} \triangleq \texttt{SAAS}(\mathbf{Q}_t, \mathbf{K}_t, \mathbf{Z}_t) \triangleq \texttt{softmax}\left(\mathbf{Q}_t (\mathbf{K}_t)^\top\right) \mathbf{Z}_t \mathbf{W}_O, \tag{5}$$

where the parameter matrix $\mathbf{W}_O \in \mathbb{R}^{N_{\text{ref}} \times N_{\mathbf{L}}}$. The learned representations from the SAAS are fed to the hierarchical reasoning and visual grounding in Section V-A.

V. Hierarchical Reasoning & Visual Grounding

A. Reasoning & Grounding Designs

The goal of our hierarchical reasoning is to capture the spatial relationship between the OoIs concerned in the HVC scenes, and overcome the challenge (2) in Section I.

Specifically, VIGOR establishes the cross-modality inter-dependencies across the input modalities for accurate visual grounding. Instead of regressing only the bounding box coordinates related to the OoIs, VIGOR aims to jointly perform the following three tasks:

1) object-level reasoning that aligns the object-level POS tags with grounded objects;
2) context-level reasoning that aligns the context-level POS tags with grounded objects; and
3) cross-modality reasoning that aligns across visual, textual, and situational modalities.

The relative positions of the OoIs can thus be differentiated, and the exact OoIs from the natural language commands can be mapped to the features returned from the heterogeneous modalities.

We note that in order to learn and capture both the object-level POS tags and the context-level POS tags, our SAAS mechanism is performed in two parallel operations, each SAAS block of which is associated with the object-level reasoning or context-level reasoning. This helps align the situational modality with the object-level and context-level POS tags hierarchically. Based on Eq. (5), these two operations of SAAS return the representations $\mathbf{K}^{(\mathbf{L}_t)}_{(\text{obj})}$ (object-level) and $\mathbf{K}^{(\mathbf{L}_t)}_{(\text{cxt})}$ (context-level), respectively. These two parallel operations of situational attention result in two different sets of parameters (i.e., $\mathbf{W}^{(\text{query})}$, $\mathbf{W}^{(\text{key})}$, $\mathbf{W}^{(\text{value})}$, and $\mathbf{W}_O$) in their blocks due to the two different levels of reasoning.

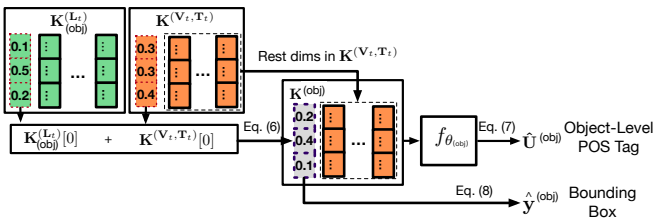

Fig. 4: Illustration of our proposed object-level reasoning.

1) Object-level Reasoning: Taking the object-level reasoning (illustrated in Fig. 4) as an example, we present our core designs as follows. Our object-level reasoning aims to learn and capture the OoIs, including the traffic participants (e.g., pedestrians, other vehicles) and infrastructures (e.g., traffic lights), relevant to the visual grounding. Recall that VIGOR extracts a sequence of POS tags from the human natural language command $\mathbf{T}_t$ that are relevant to the OoIs. VIGOR provides a multi-task learning module to jointly estimate both the object-level POS tags and the bounding boxes.

Specifically, we first take the element-wise average of the first tokens (both in $\mathbb{R}^{N_{\mathbf{T}}}$) from both $\mathbf{K}^{(\mathbf{V}_t,\mathbf{T}_t)}$ and $\mathbf{K}^{(\mathbf{L}_t)}_{(\text{obj})}$, i.e.,

$$\mathbf{K}^{(\text{obj})}_t = \frac{1}{2}\left(\mathbf{K}^{(\mathbf{V}_t,\mathbf{T}_t)}[0] + \mathbf{K}^{(\mathbf{L}_t)}_{(\text{obj})}[0]\right), \tag{6}$$

which helps focus on the most essential information respectively captured by the VTA and the SAAS. We concatenate $\mathbf{K}^{(\text{obj})}_t$ with part of $\mathbf{K}^{(\mathbf{V}_t,\mathbf{T}_t)}$, and pass them through a linear projection layer, i.e.,

$$\widehat{\mathbf{U}}^{(\text{obj})}_t \triangleq f_{\theta_{(\text{obj})}}\left(\left[\mathbf{K}^{(\text{obj})}_t;\ \mathbf{K}^{(\mathbf{V}_t,\mathbf{T}_t)}\left[1 : N^{(\mathbf{T})}_{\text{tok}} - 1\right]\right]\right), \tag{7}$$

where $\mathbf{K}^{(\mathbf{V}_t,\mathbf{T}_t)}\left[1 : N^{(\mathbf{T})}_{\text{tok}} - 1\right]$ represents the rest of the elements except for the first one. We then obtain the estimated object-level POS tags $\widehat{\mathbf{U}}^{(\text{obj})}_t \in \mathbb{R}^{N^{(\mathbf{T})}_{\text{tok}} \times N_{(\text{obj})}}$.

In the meantime, we feed $\mathbf{K}^{(\text{obj})}_t$ through another linear projection and obtain the estimated bounding box, denoted as $\widehat{\mathbf{y}}^{(\text{obj})}_t \in \mathbb{R}^4$, i.e.,

$$\widehat{\mathbf{y}}^{(\text{obj})}_t = \mathbf{W}^{(\text{obj})}\mathbf{K}^{(\text{obj})}_t, \tag{8}$$

where $\mathbf{W}^{(\text{obj})} \in \mathbb{R}^{4\times N_{\mathbf{T}}}$ represents the weight matrix to be learned. To summarize, the object-level reasoning jointly estimates object-level POS tags $\widehat{\mathbf{U}}^{(\text{obj})}_t \in \mathbb{R}^{N^{(\mathbf{T})}_{\text{tok}} \times N_{(\text{obj})}}$, where $N_{(\text{obj})}$ represents the number of object-level classes, and the bounding box $\widehat{\mathbf{y}}^{(\text{obj})}_t \in \mathbb{R}^4$ in a multi-task learning manner, as illustrated in Fig. 4.

2) Context-level Reasoning: We have the similar formulation of context-level reasoning, while the notation "obj" is replaced by "cxt", i.e., from $\mathbf{K}^{(\text{obj})}_t, \mathbf{K}^{(\mathbf{L}_t)}_{(\text{obj})}, \widehat{\mathbf{U}}^{(\text{obj})}_t$, and $\mathbf{W}^{(\text{obj})}$ to $\mathbf{K}^{(\text{cxt})}_t, \mathbf{K}^{(\mathbf{L}_t)}_{(\text{cxt})}, \widehat{\mathbf{U}}^{(\text{cxt})}_t$, and $\mathbf{W}^{(\text{cxt})}$. The context-level reasoning jointly estimates the context-level POS tags $\widehat{\mathbf{U}}^{(\text{cxt})}_t \in \mathbb{R}^{N^{(\mathbf{T})}_{\text{tok}} \times N_{(\text{cxt})}}$, where $N_{(\text{cxt})}$ represents the number of context-level classes, and the bounding box $\widehat{\mathbf{y}}^{(\text{cxt})}_t \in \mathbb{R}^4$ through another parallel pipeline of multi-task learning.

3) Cross-modality Reasoning: In addition to the above reasoning, we have designed a cross-modality reasoning module for visual grounding based on the first token returned from the tokenized embeddings from VTA, denoted as $\mathbf{K}^{(\mathbf{V}_t,\mathbf{T}_t)}[0]$, as well as two tokenized embeddings from SAAS, denoted as $\mathbf{K}^{(\mathbf{L}_t)}_{(\text{obj})}[0]$ and $\mathbf{K}^{(\mathbf{L}_t)}_{(\text{cxt})}[0]$, i.e.,

$$\mathbf{K}^{(\text{cross})}_t = \frac{1}{3}\left(\mathbf{K}^{(\mathbf{V}_t,\mathbf{T}_t)}[0] + \mathbf{K}^{(\mathbf{L}_t)}_{(\text{obj})}[0] + \mathbf{K}^{(\mathbf{L}_t)}_{(\text{cxt})}[0]\right) \in \mathbb{R}^{N_{\mathbf{L}}}. \tag{9}$$

Then, we have the estimated bounding box, denoted as $\widehat{\mathbf{y}}^{(\text{cross})}_t \in \mathbb{R}^4$, through this cross-modality grounding module, as

$$\widehat{\mathbf{y}}^{(\text{cross})}_t = \mathbf{W}^{(\text{cross})}\mathbf{K}^{(\text{cross})}_t, \tag{10}$$

where $\mathbf{W}^{(\text{cross})} \in \mathbb{R}^{4\times N_{\mathbf{L}}}$ represents the parameter matrix.

B. Model Training

To enable VIGOR to comprehend and reason the objects and contexts within the HVC scenes, we have designed the following loss functions for our model training.

• **Losses for Visual Grounding.** In order to ground the OoIs within the visual modality of HVC scenes, we jointly minimize the l-1 distances (denoted as L1) as well as the Generalized Intersections of Union (denoted as GIoU):

$$\begin{aligned} l^{(\text{obj})}_t &= \texttt{L1}\left(\mathbf{y}_t, \widehat{\mathbf{y}}^{(\text{obj})}_t\right) + \texttt{GIoU}\left(\mathbf{y}_t, \widehat{\mathbf{y}}^{(\text{obj})}_t\right) \\ &= \left|\mathbf{y}_t - \widehat{\mathbf{y}}^{(\text{obj})}_t\right| + \left(\frac{\left|\mathbf{y}_t \cap \widehat{\mathbf{y}}^{(\text{obj})}_t\right|}{\mathbf{y}_t \cup \widehat{\mathbf{y}}^{(\text{obj})}_t} - \frac{\left|\mathbf{C} \setminus \left(\mathbf{y}_t \cup \widehat{\mathbf{y}}^{(\text{obj})}_t\right)\right|}{|\mathbf{C}|}\right), \end{aligned} \tag{11}$$

where $\mathbf{C}$ represents a set of pixels of the smallest enclosing convex object for the ground-truth and estimated bounding boxes of $\mathbf{y}_t$ and $\widehat{\mathbf{y}}^{(\text{obj})}_t$. Similarly, given $\widehat{\mathbf{y}}^{(\text{cxt})}_t$ and $\widehat{\mathbf{y}}^{(\text{cross})}_t$, we find the loss functions (L1 and GIoU) of visual grounding results for context-level reasoning and cross-modality reasoning, i.e., $l^{(\text{cxt})}_t$ and $l^{(\text{cross})}_t$.

In particular, the L1 loss provides a smooth and steady training for the visual grounding, while the GIoU loss helps capture the intersection between the predicted and ground-truth bounding boxes. Combination of both helps balance between the training stability and grounding precision.

• **Losses for Object-Level & Context-Level POS Tags.** Note that both the object-level reasoning and context-level reasoning classify and return the labels of object-level and context-level POS tags. Therefore, we measure the loss functions in terms of cross entropy (denoted as CE) between the ground-truth and estimated labels, i.e., $c^{(\text{obj})}_t = \texttt{CE}(\mathbf{U}^{(\text{obj})}_t, \widehat{\mathbf{U}}^{(\text{obj})}_t)$, and $c^{(\text{cxt})}_t = \texttt{CE}(\mathbf{U}^{(\text{cxt})}_t, \widehat{\mathbf{U}}^{(\text{cxt})}_t)$.

• **Final Training Loss.** The total loss for our VIGOR training upon an HVC scene (sample) at the timestamp t is given as $l_t = \left(l^{(\text{obj})}_t + c^{(\text{obj})}_t\right) + \left(l^{(\text{cxt})}_t + c^{(\text{cxt})}_t\right) + l^{(\text{cross})}_t$.

VI. Experimental Studies

A. Experimental Settings

• **Comparison Studies.** We have compared VIGOR with the following baseline and SOTA approaches: (1) full training (FT; trained from scratch): which includes AttnGrounder [12], TransVG [6], ReSC [9], CMSVG [7], X-VLM [18], VL-BERT [5], MCN [27], SimREC [28], and D-MDETR [29]; (2) supervised fine-tuning (SFT; fine-tuned from a pre-trained model): which includes Qwen2.5-VL [30], InternVL3 [31], and MiniCPM-2.6 [32]; (3) zero-shot (ZS): which includes UniVG-R1 [33] and GPT-4o-mini. We note that our light-weight backbones are motivated by the low-latency requirement of target HVC scenarios.

• **Model Parameter & Experimental Settings.** In terms of situational modality, we follow the parameter settings in PillarNet [25]. In terms of textual modality, we follow the parameter settings (based on the first six layers) in BERT. Detailed settings are listed in the above table.

• **Experimental Environments & Evaluation Metrics.** We implement VIGOR on an HPC server with Linux Ubuntu

Symbol	Value	Symbol	Value	Symbol	Value	Symbol	Value
(W, H)	$(384, 384)$	C	3	Patch Size	32	N_{patch}	144
$N_{\text{tok}}^{(\mathbf{V})}$	145	$N_{\mathbf{V}}$	768	Max Voxel	20	$N_{(\text{voxel})}$	3×10^4
$N_{\text{tok}}^{(\mathbf{T})}$	40	$N_{\mathbf{T}}$	768	$N_{\text{tok}}^{(\mathbf{L})}$	16	$N_{\mathbf{L}}$	384
N_{ref}	64	$N_{(\text{obj})}$	3	$N_{(\text{cxt})}$	4	Learning Rate	$3\text{e}{-5}$
Batch Size	16	Epochs	120	Optimizer	AdamW	Weight Decay	$1\text{e}{-2}$

18.04.5 LTS (python 3.9.17 and pyTorch 2.1.2), AMD Ryzen 3960X, four Nvidia RTX3090 GPUs with GDDR5 24GB, and 128GB RAM. The training and testing time per HVC scene is 64ms and 41ms upon our system configuration, respectively. The following performance metrics are adopted: (i) AP@50, which represents the percentage (%) of all tested HVC scenes (samples) whose intersections of union (IoU) between the ground-truth and estimation are above 50%; (ii) mIoU, which represents the mean of IoUs (in %) among all the HVC scenes tested; and (iii) GFLOPs (Giga Floating-point Operations), which characterizes the number of floating-point operations needed (lower GFLOPs implies lighter computation) for model inference.

B. Experimental Results

Method	Type	Overall		Low-Visibility		GFLOPs↓
		AP@50↑	mIoU↑	AP@50↑	mIoU↑	
AttnGrounder [12]	FT	64.9	48.7	57.4	46.2	40.29
TransVG [6]	FT	51.2	43.2	47.9	41.9	40.48
ReSC [9]	FT	53.7	43.2	47.9	38.7	20.18
CMSVG-8 [7]	FT	62.4	52.0	56.5	47.3	17.69
CMSVG-16	FT	65.7	54.5	58.9	49.0	23.27
CMSVG-32	FT	66.6	55.0	61.6	51.4	34.44
X-VLM [18]	FT	63.5	53.6	51.8	46.3	99.62
VL-BERT [5]	FT	62.8	50.9	57.9	48.4	344.45
MCN [27]	FT	54.9	45.3	55.3	46.5	84.48
SimREC [28]	FT	58.7	48.6	59.5	50.3	20.68
D-MDETR [29]	FT	64.4	53.2	58.7	49.3	147.02
QWen2.5-VL [30]	SFT	58.0	47.9	48.8	42.4	2,708.12
InternVL3 [31]	SFT	2.5	6.4	1.5	4.5	1,050.38
MiniCPM-2.6 [32]	SFT	3.7	6.9	2.0	4.5	1,664.55
UniVG-R1 [33]	ZS	1.2	5.4	0.5	4.0	4,701.12
GPT-4o-mini	ZS	1.0	5.4	0.3	3.9	-
VIGOR	FT	**70.4**	**58.5**	**64.9**	**54.7**	99.64

• **Overall Performance.** We first show the overall performance of VIGOR and other baseline approaches.

AP@50 & mIoU. We can see that on average VIGOR outperforms the other baseline and state-of-the-art approaches by 13.6% in terms of AP@50 and 14.8% in terms of mIoU (compared with FT). We note that TransVG and AttnGrounder fall in the category of the one-stage grounding approaches, and may not necessarily capture the OoIs in our complex traffic environments. ReSC adopts the long short-term memory encoder, which, however, may not capture the fine-grained textual features in the HVC scenes. While CMSVG (underlined) may achieve overall comparable performance with VIGOR when given many textual proposals of the objects [7], the proposal-demanding property (and subsequent needs of large memory in processing the proposals) may not be feasible for CMSVG to be efficiently trained and deployed upon the resource-constrained on-board computers. We have also evaluated, for instance, X-VLM fused with situational modality (LiDAR), whose AP@50 drops from 63.5% to 59.5%, implying importance of a proper heterogeneous representation learning design as ours.

For the SFT models, we observed that performance is highly correlated with the base models. QWen is specially optimized for the grounding task, resulting in relatively good performance above others. The low performance of the ZS models indicates that these models may struggle when the dataset deviates from its normal pre-training distribution and the language expression is ambiguous.

Low-Visibility. We have showcased all approaches upon a total of 675 low-visibility scenes (e.g., nighttime and rainy conditions). We evaluate VIGOR in grounding the OoIs under the low-visibility scenes. VIGOR outperforms the other FT baseline approaches in AP@50 and mIoU metrics by 14.1% and 14.4%, respectively. Our results demonstrate VIGOR's effectiveness in grounding the OoIs under the low-visibility HVC scenes, particularly thanks to the integration of situational modalities.

Efficiency. Compared with vision-language models like QWen2.5-VL and InternVL3, VIGOR, similar to other FT-based approaches exhibit much smaller computational requirements. VIGOR also demonstrates a strong balance between model performance and computational efficiency.

• **Model Ablation Studies.** We have also conducted the model ablation studies by excluding the modules or tasks within VIGOR. We compare VIGOR with (1) without (w/o) situational modality, (2) w/o hierarchical reasoning, (3) w/o $\widehat{y}_t^{(\text{obj})}$ in the object-level reasoning, (4) w/o $\widehat{y}_t^{(\text{cxt})}$ in the context-level reasoning, (5) w/o $\widehat{y}_t^{(\text{cross})}$ in Eq. (10), (6) w/o adaptive sampling in Eq. (2). We can observe that all variations experience the performance drop from the full version of VIGOR. From the performance drops in (2) and (3) from the full version, we can particularly infer the importance of the hierarchical reasoning and the object-level estimation. While (1) w/o situational modality already outperforms all the visual grounding baselines (see table in Overall Performance), the performance drop from the full version to (1) also demonstrates the importance of additional situation awareness (LiDAR) in discerning the OoIs.

Type	VIGOR (full)	(1)	(2)	(3)	(4)	(5)	(6)
AP@50	70.4	68.8	68.4	68.4	68.6	69.0	69.2

• **Case Studies, Visualization, & Comparison.** Fig. 5 shows the visualization results of VIGOR (highlighted in red) in comparison with VL-BERT [5] (highlighted in green). We have further showcased four sets of HVC scenes with vehicles, pedestrians, infrastructures, and low-light/low-visibility conditions. We also attach per each HVC scene the corresponding the human natural language command with the object-/context-level POS tags.

One can see that VIGOR outperforms VL-BERT in grounding the OoIs. Taking the "vehicle" (the first row in Fig. 5) as an example, we have shown that VIGOR yields accurate visual grounding in the HVC scenes of parking and following. We can observe that VL-BERT may only capture the key OoIs "car" without reasoning about the actual one from other vehicles. On the other hand, VIGOR provides both the object-level and context-level reasoning, thus yielding more accurate visual grounding results.

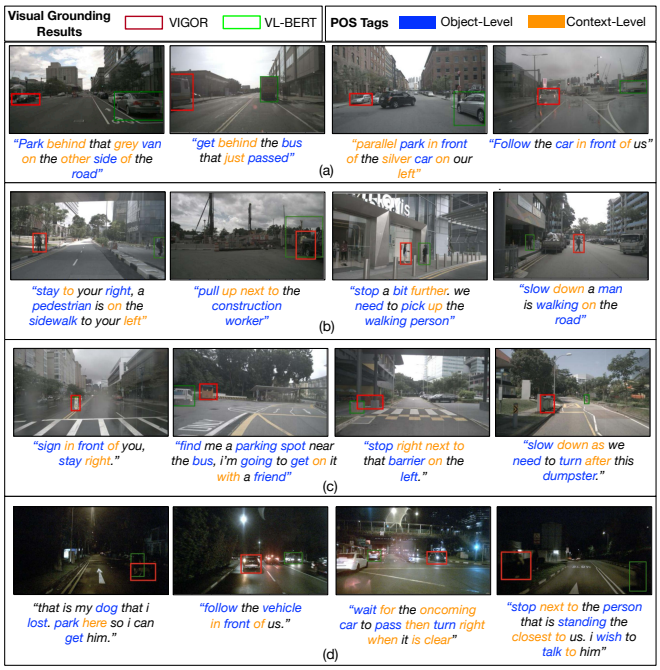

Fig. 5: Examples in various HVC scenes, i.e., (a) vehicles, (b) pedestrians, (c) infrastructures, and (d) low-light/low-visibility.

VII. Conclusion

We have designed and evaluated VIGOR, a novel visual grounding approach based on cross-modality representation learning and hierarchical reasoning for human-to-vehicle commanding (HVC) scenes in complex traffic environments. Taking into account the visual, textual, and situational modalities, we have designed within VIGOR the heterogeneous representation learning task, as well as the visual grounding and hierarchical reasoning task. Our evaluation results over various HVC scenes have corroborated the effectiveness and accuracy of VIGOR in grounding the OoIs, and overcoming the challenges in terms of complex traffic environments and reasoning the OoIs for their differentiation.

Acknowledgment

This project is supported, in part, by the National Science Foundation (NSF) under Grants No. 2239897, 2303575, 2245223, Google Research Scholar Program Award, and NVIDIA Applied Research Accelerator Program Award.

References

[1] M. Tabatabaie, S. He, *et al.*, "Beyond "taming electric scooters": Disentangling understandings of micromobility naturalistic riding," *Proc. ACM IMWUT*, vol. 8, no. 3, Sept. 2024.

[2] H. Wang and S. He, "Vision-Language Modeling for Scene Understanding and Reasoning of Vehicle-to-X Interactions," in *Proc. IEEE MASS*, 2025, pp. 174–182.

[3] M. Tabatabaie, S. He, *et al.*, "Interaction-Aware and Hierarchically-Explainable Heterogeneous Graph-based Imitation Learning for Autonomous Driving Simulation," in *Proc. IEEE/RSJ IROS*, 2023, pp. 3576–3581.

[4] S. Kuutti, R. Bowden, *et al.*, "A survey of deep learning applications to autonomous vehicle control," *IEEE T-ITS*, vol. 22, no. 2, pp. 712–733, 2020.

[5] W. Su, X. Zhu, *et al.*, "VL-BERT: Pre-training of generic visual-linguistic representations," *arXiv preprint arXiv:1908.08530*, 2019.

[6] J. Deng, Z. Yang, *et al.*, "TransVG: End-to-end visual grounding with transformers," in *Proc. ICCV*, 2021, pp. 1769–1779.

[7] N. Rufus, U. K. R. Nair, *et al.*, "Cosine meets softmax: A tough-to-beat baseline for visual grounding," in *Proc. ECCV Workshops*. Springer, 2020, pp. 39–50.

[8] P. Schafhalter, S. Kalra, *et al.*, "Leveraging Cloud Computing to Make Autonomous Vehicles Safer," in *Proc. IEEE/RSJ IROS*, 2023, pp. 5559–5566.

[9] Z. Yang, T. Chen, *et al.*, "Improving one-stage visual grounding by recursive sub-query construction," in *Proc. ECCV*. Springer, 2020, pp. 387–404.

[10] H. Caesar, V. Bankiti, *et al.*, "nuScenes: A multimodal dataset for autonomous driving," in *Proc. CVPR*, 2020, pp. 11 621–11 631.

[11] T. Deruyttere, S. Vandenhende, *et al.*, "Talk2Car: Taking control of your self-driving car," *arXiv preprint arXiv:1909.10838*, 2019.

[12] V. Mittal, "AttnGrounder: Talking to cars with attention," in *Proc. ECCV Workshops*. Springer, 2020, pp. 62–73.

[13] D. Grujicic, T. Deruyttere, *et al.*, "Predicting physical world destinations for commands given to self-driving cars," in *Proc. AAAI*, vol. 36, no. 1, 2022, pp. 715–725.

[14] W. Cheng, J. Yin, *et al.*, "Language-guided 3d object detection in point cloud for autonomous driving," *arXiv preprint arXiv:2305.15765*, 2023.

[15] A. B. Vasudevan, D. Dai, *et al.*, "Talk2Nav: Long-range vision-and-language navigation with dual attention and spatial memory," *International Journal of Computer Vision*, vol. 129, pp. 246–266, 2021.

[16] A. Waswani, N. Shazeer, *et al.*, "Attention is all you need," in *Proc. NeurIPS*, 2017.

[17] Y. Qiao, C. Deng, *et al.*, "Referring expression comprehension: A survey of methods and datasets," *IEEE TMM*, vol. 23, pp. 4426–4440, 2020.

[18] Y. Zeng, X. Zhang, *et al.*, "Multi-Grained Vision Language Pre-Training: Aligning Texts with Visual Concepts," *arXiv preprint arXiv:2111.08276*, 2021.

[19] L. Xiao, X. Yang, *et al.*, "HiVG: Hierarchical multimodal fine-grained modulation for visual grounding," in *Proc. ACM MM*, 2024, pp. 5460–5469.

[20] M. Xie, M. Wang, *et al.*, "Phrase decoupling cross-modal hierarchical matching and progressive position correction for visual grounding," *IEEE Transactions on Multimedia*, 2025.

[21] M. Dai, L. Yang, *et al.*, "SimVG: A simple framework for visual grounding with decoupled multi-modal fusion," *Proc. NeurIPS*, vol. 37, pp. 121 670–121 698, 2024.

[22] M. Dai, J. Li, *et al.*, "Multi-task visual grounding with coarse-to-fine consistency constraints," in *Proc. AAAI*, vol. 39, no. 3, 2025, pp. 2618–2626.

[23] L. Yang, Y. Xu, *et al.*, "Improving visual grounding with visual-linguistic verification and iterative reasoning," in *Proc. CVPR*, 2022, pp. 9499–9508.

[24] Z. Liu, Y. Lin, *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF ICCV*, 2021, pp. 10 012–10 022.

[25] G. Shi, R. Li, *et al.*, "PillarNet: Real-time and high-performance pillar-based 3d object detection," in *Proc. ECCV*. Springer, 2022, pp. 35–52.

[26] J. Dai, H. Qi, *et al.*, "Deformable convolutional networks," in *Proc. IEEE ICCV*, 2017, pp. 764–773.

[27] G. Luo, Y. Zhou, *et al.*, "Multi-task collaborative network for joint referring expression comprehension and segmentation," in *Proc. CVPR*, 2020, pp. 10 034–10 043.

[28] ——, "A Survivor in the Era of Large-Scale Pretraining: An Empirical Study of One-Stage Referring Expression Comprehension," *IEEE TMM*, 2023.

[29] F. Shi, R. Gao, *et al.*, "Dynamic MDETR: A Dynamic Multimodal Transformer Decoder for Visual Grounding," *IEEE. TPAMI*, vol. 46, no. 2, pp. 1181–1198, 2024.

[30] S. Bai, K. Chen, *et al.*, "Qwen2.5-VL Technical Report," *arXiv preprint arXiv:2502.13923*, 2025.

[31] Z. Chen, J. Wu, *et al.*, "InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks," in *Proc. CVPR*, 2024, pp. 24 185–24 198.

[32] Y. Yao, T. Yu, *et al.*, "MiniCPM-V: A GPT-4V Level MLLM on Your Phone," *arXiv preprint arXiv:2408.01800*, 2024.

[33] S. Bai, M. Li, *et al.*, "UniVG-R1: Reasoning Guided Universal Visual Grounding with Reinforcement Learning," *arXiv preprint arXiv:2505.14231*, 2025.