Highlights

GPT-4 Enhanced Multimodal Grounding for Autonomous Driving: Leveraging Cross-Modal Attention with Large Language Models

Haicheng Liao,Huanming Shen,Zhenning Li,Chengyue Wang,Guofa Li,Yiming Bie,Chengzhong Xu

- Developing a pioneering hybrid strategy for advanced image-text context analysis in autonomous vehicle command grounding.
- Introducing a novel cross-modal attention mechanism, uniquely tailored to derive insightful human-AV interaction from multi-modal inputs.
- Leveraging the sophisticated capabilities of the large language model GPT-4 for effective embedding and nuanced interpretation of human emotions in commander commands.
- Demonstrating exceptional robustness and adaptability of our model across a spectrum of challenging traffic scenarios, validated extensively on the Talk2Car dataset.

GPT-4 Enhanced Multimodal Grounding for Autonomous Driving: Leveraging Cross-Modal Attention with Large Language Models

Haicheng Liao[a,1], Huanming Shen[b,1], Zhenning Li[c,*], Chengyue Wang[d], Guofa Li[e], Yiming Bie[f] and Chengzhong Xu[a,*]

[a]State Key Laboratory of Internet of Things for Smart City and Department of Computer and Information Science, University of Macau, Macau SAR, China

[b]Department of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu, China

[c]State Key Laboratory of Internet of Things for Smart City and Departments of Civil and Environmental Engineering and Computer and Information Science, University of Macau, Macau SAR, China

[d]State Key Laboratory of Internet of Things for Smart City and Departments of Civil and Environmental Engineering, University of Macau, Macau SAR, China

[e]College of Mechanical and Vehicle Engineering, Chongqing University, Chongqing, China

[f]School of Transportation, Jilin University, Changchun, China

ARTICLE INFO

Keywords:
Autonomous Driving
Visual Grounding
Cross-Modal Attention
Large Language Models
Human-Machine Interaction

ABSTRACT

In the field of autonomous vehicles (AVs), accurately discerning commander intent and executing linguistic commands within a visual context presents a significant challenge. This paper introduces a sophisticated encoder-decoder framework, developed to address visual grounding in AVs. Our Context-Aware Visual Grounding (CAVG) model is an advanced system that integrates five core encoders—Text, Image, Context, and Cross-Modal—with a Multimodal decoder. This integration enables the CAVG model to adeptly capture contextual semantics and to learn human emotional features, augmented by state-of-the-art Large Language Models (LLMs) including GPT-4. The architecture of CAVG is reinforced by the implementation of multi-head cross-modal attention mechanisms and a Region-Specific Dynamic (RSD) layer for attention modulation. This architectural design enables the model to efficiently process and interpret a range of cross-modal inputs, yielding a comprehensive understanding of the correlation between verbal commands and corresponding visual scenes. Empirical evaluations on the Talk2Car dataset, a real-world benchmark, demonstrate that CAVG establishes new standards in prediction accuracy and operational efficiency. Notably, the model exhibits exceptional performance even with limited training data, ranging from 50% to 75% of the full dataset. This feature highlights its effectiveness and potential for deployment in practical AV applications. Moreover, CAVG has shown remarkable robustness and adaptability in challenging scenarios, including long-text command interpretation, low-light conditions, ambiguous command contexts, inclement weather conditions, and densely populated urban environments. The code for the proposed model is available at our Github.

1. Introduction

While autonomous vehicles (AVs) herald a transformative era in urban mobility and have the potential to significantly reduce road accidents, their widespread acceptance remains hampered by public skepticism (Li et al., 2023a). The root of this hesitancy often lies in concerns about trust and the perceived loss of control. These issues are exacerbated when AVs display indecisiveness or ambiguity in complex driving scenarios (Li et al., 2022a). To address this trust gap, our research introduces a pioneering framework that augments human-AV interaction by effectively grounding natural language commands within visual contexts. Envision a scenario where a commander instructs an AV with a simple command like "Park in a shaded area on a sunny day." Such interactions not only empower the commander but also infuse the vehicle's decision-making process with essential context, especially in ambiguous or complex driving situations. Recent studies further highlight the significant enhancement in commander experience and acceptance when natural language commands facilitate clear communication of intentions to AVs (Deruyttere et al., 2021; Othman, 2021; Bonnefon et al., 2016; Dong et al., 2023a).

*Corresponding author; [1]Equally Contributed
zhenningli@um.edu.mo (Z. Li); czxu@um.edu.mo (C. Xu)



Command1: Oh this guy on the street is my boss, actually but I called in sick. Turn left soon!
Emotion classification: Urgent



Command2: Wow hold on! That looks like my stolen bike over there! Drop me off next to it.
Emotion classification: Urgent

Figure 1: Illustration of Regions Identified by an AV based on a Raw Image and a Natural Language Command. The blue bounding box represents the ground truth. The red and yellow bounding boxes correspond to the prediction results from CAVG with emotion categorization and without emotion categorization, respectively.

However, decoding commander intent poses a substantial challenge for AVs, primarily due to the intricacies of interactions between commanders and AVs. To effectively comprehend and act upon these commands, AVs must adeptly capture the complex semantic relationship between natural language and the dynamic visual environment. This task is further complicated by the multifaceted nature of real-world traffic scenes, characterized by ever-changing visual elements such as traffic agents, lighting conditions, road markings, and fluctuating weather. These elements, coupled with the inherent ambiguity of natural language commands, can lead to potentially hazardous misunderstandings between commanders and AVs, resulting in unsafe driving maneuvers.

Addressing these challenges necessitates the development of robust models capable of interpreting and grounding natural language commands amidst this complex and continuously evolving visual landscape. Moreover, considering the emotional and psychological dimensions of human-AV interactions is critical (Tang et al., 2015; Liao et al., 2023). Natural language commands inherently carry a spectrum of emotional overtones—from informative and commanding to urgent. Neglecting the emotional aspect and its fluctuating responses in different traffic contexts could hinder the ability to discern subtle variations in commands. These emotional variables, intertwined with language use, significantly influence the underlying intent and nuances conveyed in commands. Consequently, adopting a holistic approach that encompasses the emotional dimension of human-AV interaction is imperative.

Despite considerable progress in this field, with various methodologies and frameworks proposed (Mittal, 2020; Deruyttere et al., 2020a; Rufus et al., 2020), existing solutions typically rely on a segmented process of object identification followed by linguistic alignment. These methods focus primarily on the bounding boxes of detected objects, often overlooking essential real-world driving contexts. Furthermore, traditional approaches usually concentrate on surface-level features, neglecting underlying elements that could enrich the interpretation and execution of commands. In contrast, recent advancements (Bhattacharyya et al., 2022; Dong et al., 2021; Wen et al., 2021) have adopted integrated frameworks that merge textual and visual elements early in the process but face limitations in handling complex multi-object scenarios. Even with the introduction of Transformer-based models (Ding et al., 2021, 2022; Dong et al., 2022), challenges remain in navigating intricate real-world situations. Importantly, most state-of-the-art models (Grujicic et al., 2022; Chan et al., 2022; Jain et al., 2023) in language grounding tend to overlook the emotional dynamics inherent in human-AV interactions, a crucial aspect that our research aims to address.

Our model introduces a novel approach that fuses advancements in visual-language tasks and image-text retrieval. Adopting a sophisticated, two-pronged strategy, it capitalizes on the strengths of both single-stream and dual-stream architectures. The first strategy enhances contextual understanding by integrating contextual cues into the CAVG model, elevating its ability to interpret and respond to natural language commands in real-world driving scenarios. The second strategy involves implementing a multimodal prediction framework, allowing the system to anticipate various plausible outcomes, each paired with a calculated likelihood. This approach is particularly beneficial in mitigating risks and uncertainties in complex and ambiguous driving situations. Additionally, CAVG categorizes subtle emotions inherent in human language, shaping its responses to align with the psychological and emotional context presented by the commanders. As illustrated in Fig. 1, CAVG excels at grounding semantically matched objects in dynamic

visual traffic scenes in response to natural language commands, adapting its responses to the emotional context of the commander while maintaining contextual accuracy.

In conclusion, our research introduces a novel model that revolutionizes the way commanders interact with AVs through natural language commands. This innovative approach promises to transform the landscape of human-AV interaction significantly, enhancing commander comfort, enriching user experience, and elevating satisfaction levels. Our model's distinct contributions are as follows:

- We have developed a unique approach utilizing a pre-trained model to blend semantic information with global image features. This results in a cross-modal vector, which forms the basis of our novel multi-head cross-model attention mechanism. This mechanism adeptly integrates global context and textual information, enriching the understanding of each sub-region in the visual field.
- Our model incorporates an innovative emotion analysis component using Generative Pretrained Transformer 4 (GPT-4). This allows us to capture the nuanced emotional content in natural language commands. This feature makes the interaction between humans and AVs more intuitive and responsive, enabling the vehicle to understand and react to the commander's intent in a more human-centric manner.
- The Cross-Modal Encoder in our model facilitates the extraction of interactive features between different input types, such as images and commands. Utilizing a Region-Specific Dynamic (RSD) layer attention mechanism, our model adeptly analyzes and prioritizes multi-modal data, offering a more refined understanding of the environment.
- Our empirical evaluation demonstrates that our model surpasses existing state-of-the-art methods in performance, particularly on the challenging Talk2Car dataset. This includes scenarios like low-light conditions, long-text commands, rainy weather, and dense urban environments. Our model achieves this with less training time, superior performance, and enhanced interpretability.

This paper is structured in the following manner to facilitate a comprehensive understanding: Section 2 offers a detailed review of the existing literature and studies pertinent to our research area. In Section 3, we articulate the research task at hand and present a detailed diagram of our model's architecture. Section 4 elaborates on the methodology of our experimental setup, followed by an in-depth discussion of the results obtained. In Section 5, we investigate the utility and user satisfaction of our proposed model using a questionnaire. Finally, Section 6 synthesizes the main findings of this study and proposes potential directions for future investigations in this field.

2. Related Work

2.1. Visual Grounding

Visual Grounding (VG) is a task that fundamentally involves localizing image regions pertinent to a specific natural language command. The body of research in this domain typically falls into two primary categories based on their methodological frameworks: one-stage and two-stage methods.

One-Stage Methods in Visual Grounding: These methods are characterized by their efficiency, primarily due to a streamlined processing paradigm. Unlike the two-stage models, which initiate with proposing candidate regions followed by object classification refinement, one-stage methods (Chen et al., 2018; Yang et al., 2019b; Liao et al., 2020; Yang et al., 2020, 2022) simultaneously extract image and command features in a unified pass, predicting both object categories and their bounding boxes concurrently. For instance, Chen et al. (2018) introduced a guided attention mechanism to focus on the referent's central region, interpreting referring expressions in a single stage without resorting to region proposals. Liao et al. (2020) reformulated referring expression comprehension as a correlation filtering process for precise object localization. Moreover, Yang et al. (2022) integrated a visual-linguistic verification module within a unified one-stage framework to effectively identify regions corresponding with descriptions. The elimination of the region proposal phase in single-stage methods facilitates faster inference times but often requires extended training durations and may involve trade-offs in performance. Particularly, these methods may struggle in scenarios with densely packed objects or overlapping regions of interest, posing accuracy challenges and limiting their robust application in AV contexts.

Two-Stage Methods in Visual Grounding: In contrast, two-stage methods are renowned for their meticulous object identification process within images. These methods typically employ pre-trained models as object detectors in

the first stage to extract region-specific features (Wang et al., 2019; Tan and Bansal, 2019; Wang et al., 2018; Zhuang et al., 2018). For example, MAttNet (Yu et al., 2018) and Volta (Bugliarello et al., 2021) utilize Faster R-CNN (Ren et al., 2015) and ResNet-101 (He et al., 2016) as their backbone networks to capture region features. These features form the basis for defining the spatial locations of potential objects of interest, termed region proposals. In the second stage, these proposals are integrated with linguistic features and analyzed by the framework's decoder, which selects regions that align closely with the linguistic input, ensuring a semantically coherent and contextually appropriate detection result. While two-stage methods have demonstrated efficacy in VG tasks, they are often heavily reliant on the accuracy of pre-trained object detection models. This reliance can be a drawback, especially when issues in bounding box localization occur or when crucial semantic information outside the proposed regions is overlooked. Our study adopts a two-stage approach but extends its focus beyond traditional bounding boxes. We aim to capture a broader spectrum of contextual information within various traffic scenarios. This approach seeks to leverage the synergies between bounding box localization and contextual cues, enhancing object localization and promoting context-aware recognition in AV applications.

2.2. Language-to-Image in Autonomous Driving

The fusion of object detection with semantic information processing has catalyzed significant advancements in AVs, particularly in the realm of language grounding. A seminal work in this area is the Talk2Car dataset introduced by Deruyttere et al. (Deruyttere et al., 2019). This innovative dataset, comprising natural language commands specifically tailored for self-driving cars, serves as a critical benchmark and sheds light on the challenges of object referral in VG for AVs.

Progressing from the foundations established by the Talk2Car dataset, recent research has further advanced the field. Mittal's AttnGrounder (Mittal, 2020), for instance, is an end-to-end model with a Cross-Modal attention module that correlates words in a query to regions in visual scenes. Rufus et al. (Rufus et al., 2020) contributed a minimalist yet effective approach to VG in AVs. Further enriching this domain, Deruyttere et al. (Deruyttere et al., 2020b, 2021) and Rufus et al. (Rufus et al., 2021) explored uncertainties in commands and navigable region identification. Chan et al. (Chan et al., 2022) and Cheng et al. (Cheng et al., 2023) have also made significant strides, with Cheng et al. introducing LiDAR Grounding. Dong et al. (Dong et al., 2023b) presented the HuBo-VLM model, showcasing a unified vision-language approach for human-robot interactions.

2.3. Emotion Classification

Emotion classification has been a focus in several studies, including those analyzing emotions expressed in specific events (Li et al., 2023b, 2022c). Devlin et al.'s BERT (Devlin et al., 2018) integrated transformer encoders for pretraining models on NLP tasks, including emotion classification. Yang et al. (Yang et al., 2019a) introduced EmotionX-KU, a classifier for detecting emotions in dialogue, as part of the EmotionX 2019 challenge. Sentiment GPT (Kheiri and Karimi, 2023) applied GPT models to emotion classification, highlighting their effectiveness in sentiment analysis. In our study, we utilize GPT-4 for analyzing emotional content in commands, drawing on its extensive pre-training to grasp subtle linguistic patterns and contextual nuances.

2.4. Vision-and-Language Model

The domain of Vision-and-Language pre-trained models is characterized by two principal architectures: single-stream and dual-stream. Single-stream models like ImageBert (Qi et al., 2020) seamlessly integrate visual and linguistic modalities across encoder layers for holistic data processing. In contrast, dual-stream architectures, such as LXMERT (Tan and Bansal, 2019) and VilBERT (Su et al., 2019), begin with intra-modal interactions, gradually merging modalities as processing progresses. LXMERT, for example, employs an object-relationship encoding module to align complex visual patterns with linguistic cues. In our research, we adopt a hybrid approach that harnesses the strengths of both architectures, optimizing the CAVG model for intricate tasks in AVs.

3. Proposed Model

3.1. Overview

Our aim is to devise a method that can process a frontal-view image of a vehicle, denoted as $\boldsymbol{I}$, in tandem with a natural language command, denoted as $\boldsymbol{C}$. The ultimate objective is to accurately pinpoint the destination region within the image where the vehicle should navigate in order to comply with the given command. To formalize this, we cast

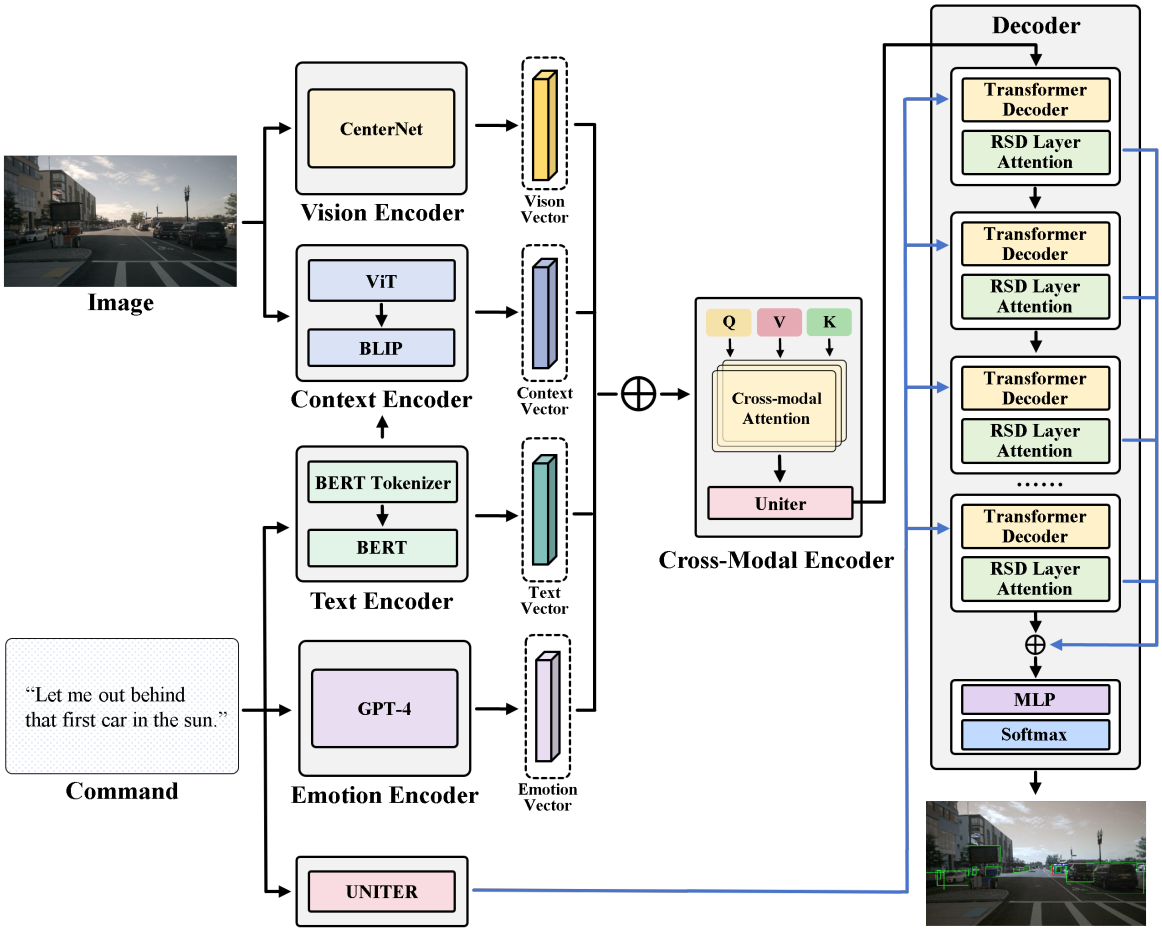


Figure 2: Schematic of the Model Architecture. The Text Encoder and the Emotion Encoder generate a text vector and an emotion vector, respectively, from the given command, while the Vision Encoder divides the input image into N RoIs, each represented by a vision vector. These vectors are contextually enriched by a context encoder and then merged by a Cross-Modal Encoder using multi-head cross-modal attention. The multimodal decoder calculates likelihood scores for each region and selects the top-k regions that best match the semantics of the command. The final prediction is based on this fusion.

the task as a mapping problem: identifying the most suitable Regions of Interest (RoIs) in the image I that corresponds to the destination the vehicle must reach based on the natural language command C. This mapping is not trivial and involves determining the best-matching bounding boxes within the image that are in line with the semantic intent of the command.

As illustrated in Fig. 2, our proposed framework utilizes an encoder-decoder architecture that comprises five specialized encoders: the Text Encoder, the Emotion Encoder, the Vision Encoder, the Context Encoder, and the Cross-Modal Encoder. These are integrated with a Multimodal Decoder to perform the task of destination identification.

3.2. Text Encoder

The Text Encoder is primarily responsible for processing the natural language command C. To achieve this, we use the well-known Bidirectional Transformer (BERT) architecture (Devlin et al., 2018). BERT has emerged as a seminal advancement in NLP due to its remarkable capabilities in capturing intricate contextual nuances and relationships within textual data. From a technical perspective, BERT uses a multi-layer bidirectional transformer architecture with a deep stack of transformer encoder layers. These layers include self-attention mechanisms that allow the model

to capture dependencies between words in both forward and backward directions, ensuring a thorough contextual understanding of the entire input sequence.

Specifically, the Text Encoder uses BERT's robust pre-trained language representations to generate a comprehensive text vector, denoted $\boldsymbol{O}_{text}$. The input command $\boldsymbol{C}$ is tokenized into a sequence using BERT's WordPieces tokenizer and then fed into the BERT model. The resulting output is a unified command embedding $\{w_1, w_2, w_3, \ldots, w_n\}$ that serves as $\boldsymbol{O}_{text}$. This embedding encapsulates the textual characteristics of the input command, including its meaning and context. It lays the foundation for subsequent multimodal fusion and contextual analysis, enabling CAVG to seamlessly integrate and process language commands in the realm of autonomous driving. Formally,

$$\boldsymbol{O}_{text} = \phi_{BERT}\left(\phi_{Tokenizer}(\boldsymbol{C})\right) = \{w_1, w_2, w_3, \ldots, w_n\} \tag{1}$$

where $\phi_{Tokenizer}(\boldsymbol{C})$ represents the transformation of the input command $\boldsymbol{C}$ into its corresponding token embeddings using the BERT's WordPiece tokenizer, while the $\phi_{BERT}(\cdot)$ denotes the BERT architecture. In particular, our Text Encoder is more than just an off-the-shelf implementation of BERT. Instead, we are fine-tuning and adapting BERT to the specific requirements and nuances of language used in the context of autonomous driving. This fine-tuning process adjusts the model to better understand and interpret the unique vocabulary, syntax, and command structures encountered in the context of autonomous driving.

3.3. Emotion Encoder

The Emotion Encoder in our study is crucial for interpreting emotions in commander commands, which is vital for the safe and efficient operation of AVs. This encoder is adept at classifying human emotions, drawing on the advanced capabilities of Large Language Models (LLMs), specifically GPT-4, as referenced by (OpenAI, 2023). GPT-4's pre-training on a broad corpus of text data equips it with a deep understanding of complex linguistic patterns, semantics, and contextual nuances. Its core self-attention mechanism enables nuanced emotion analysis, allowing it to assign different levels of importance to various words in a sentence.

The Emotion Encoder processes commander commands and categorizes them into three distinct emotional states:

- **Urgent:** These commands require immediate action due to their time-sensitive or critical nature. For example, a commander yelling "Wow hold on! That looks like my stolen bike over there! Drop me off next to it.", conveys a sense of urgency that requires immediate attention. The Emotion Encoder recognizes this emotional state and interprets this command as requiring immediate attention. Consequently, the destination is inferred as the closest and safest place to pull over, in line with the commander's anxious emotional state.

- **Commanding:** These commands are characterized by their direct and unambiguous nature, specifying what actions the vehicle should take without an inherent time sensitivity. For instance, a command like, "Make a left turn at the next intersection.", is clear and commanding. While the command itself is straightforward, our Emotion Encoder detects the absence of urgency in the commander's tone. It categorizes this as a "commanding" instruction. In response, the model processes the command accordingly, planning a left turn at the next intersection without the need for immediate action.

- **Informative:** Neutral instructions that convey useful information without a strong emotional tone. Informative commands are primarily characterized by their objective nature, focusing on providing factual details about a particular subject or object. These commands tend to be descriptive, where the content predominantly revolves around elucidating specific attributes or features of an entity without conveying a pressing need for action.

The Emotion Encoder utilizes the GPT-4 API for emotion analysis. The API setup involves secure communication with OpenAI's servers, and each commander command is preprocessed to align with GPT-4's input requirements. The command is sent to GPT-4, which then analyzes the text and returns a response indicating the emotional tone. This response is post-processed to categorize the command into one of the predefined categories.

The output of the Emotion Encoder $\boldsymbol{O}_{emo}$ is the predicted emotion category e for the input command $\boldsymbol{C}$, which can be defined as follows:

$$\boldsymbol{O}_{emo} = \phi_{GPT}(\boldsymbol{C}) = \{e\} \tag{2}$$

where the $\phi_{GPT}(\cdot)$ represents the GPT-4 framework. Subsequently, the emotion vector $\boldsymbol{O}_{emo}$ is fused with the text vector $\boldsymbol{O}_{text}$ to incorporate the emotional context. This integration is devised to indicate CAVG towards providing human-like responses that resonate with both the given command and the emotional disposition of the commander.

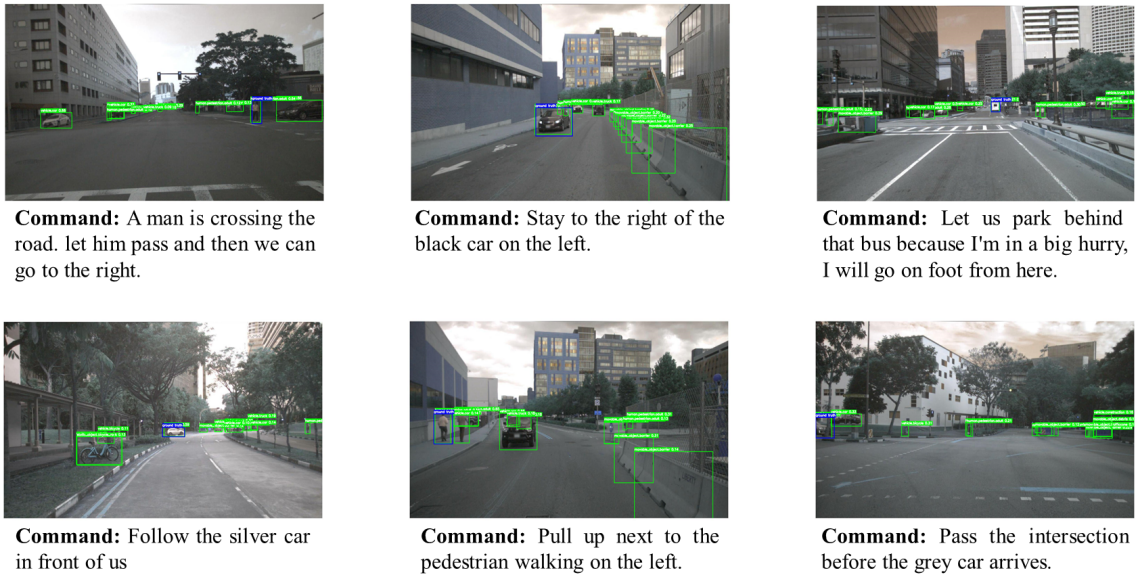


Figure 3: The Output of the Vision Encoder on the Talk2Car Dataset. Each bounding box is labeled with an attribute class followed by an object class. Ground truth bounding boxes are depicted in blue, while the output bounding boxes of CAVG are highlighted in green. A natural language command associated with each visual scenario is also displayed below the image for context.

3.4. Vision Encoder

The Vision Encoder is architected to extract rich visual information from the given view image $\boldsymbol{I}$ for subsequent multimodal analysis. Within this encoder, the CenterNet (Dosovitskiy et al., 2020) model is utilized to identify and localize entities and elements within the observed scene in the raw image. Specifically, CenterNet is built on the foundation of two basic components: Fast R-CNN (Ren et al., 2015) and ResNet-101 (He et al., 2016). Known for its efficiency and accuracy, Fast R-CNN utilizes a region proposal network (RPN) to generate region proposals. These proposals are then subjected to a two-stage process involving a classifier and a regressor. The classifier is responsible for object detection, while the regressor refines the bounding box coordinates. As shown in Fig. 3, through this process, CAVG can effectively identify prominent entities and detect their classes in the input image, such as trees, vehicles, bicycles, and pedestrians. In particular, these identified entities are encapsulated as RoIs, rich in semantic features and spatial correlations among different entities.

Complementing the capabilities of Fast R-CNN, ResNet-101 is introduced into the framework to enhance feature representation, with a keen focus on capturing finer details, textures, and distinguishing attributes of detected objects. Such nuances are essential for understanding commands that may refer to specific visual cues or unique scene features. For example, given a visual scene with the command “Could you please park before the red car?”, the ResNet-101 is adept at capturing subtle color variations and textures within the RoIs and embedding them into high-dimensional visual representations. This empowers CAVG to pinpoint the relevant vehicles and greatly helps to provide accurate, context-aware responses.

Overall, this synergistic integration within the CenterNet model enables the encoder to produce the high-dimensional visual representations, encapsulated in a vision vector, labeled $\boldsymbol{O}_{vision}$, which can be given as:

$$\boldsymbol{O}_{vision} = \phi_{CenterNet}(C) = \{b_1, b_2, b_3, \ldots, b_n\} \tag{3}$$

Here, $\phi_{CenterNet}(\cdot)$ represents the CenterNet model and $\{b_1, b_2, b_3, \ldots, b_n\}$ is the processed RoIs in the vision vector.

3.5. Context Encoder

Our model confronts a more intricate challenge than traditional Referring Image Segmentation tasks due to its multimodal nature and the unstructured nature of its ground truth. In the “talk-to-car" scenario, the fusion of

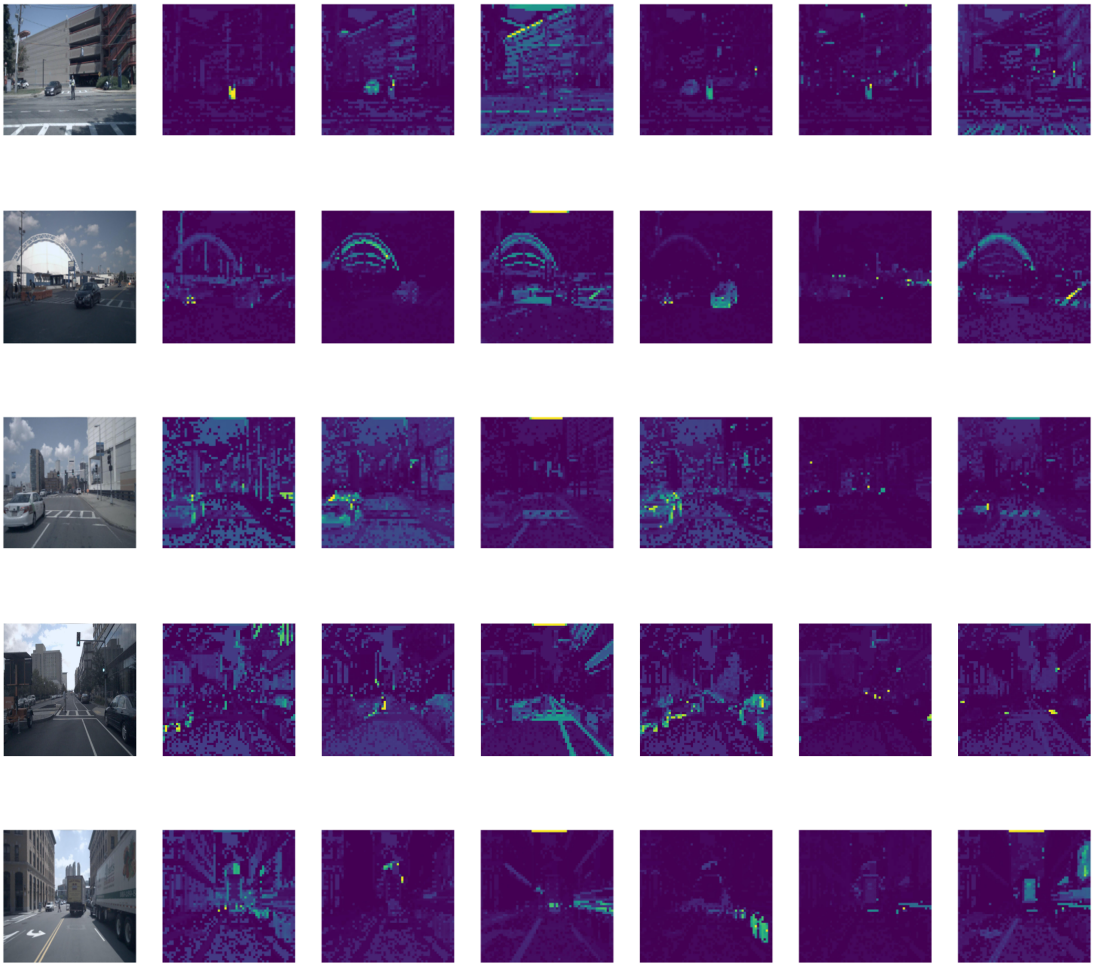


Figure 4: Raw Images and Examples Output from Context Encoder. CAVG automatically learns to perform object segmentation without a complex self-supervised training recipe or any fine-tuning with segmentation-related annotations. For each image set, we visualize the raw image on the left and the self-attention maps of the image in different layers of the encoder on the right.

unrestricted linguistic expressions with diverse visual scenes poses several challenges. These include co-reference resolution, precise identification of named entities, understanding complex object relationships, accurate association of attributes with corresponding visual elements, and pinpointing the object of interest within a command.

Diverging from existing methods that primarily focus RoIs (Mittal, 2020; Hu et al., 2016; Deruyttere et al., 2020a), our Context Encoder adopts a more holistic approach. It is designed not only to identify key focal points within the input image but also to discern the broader contextual relationships within the entire visual scene, going beyond the limitations of bounding boxes. To achieve this, we utilize the Vision Transformer (ViT) and the state-of-the-art Vision-and-Language (V&L) pre-training model, and LLM Bootstrapping Language-Image Pre-training Model (BLIP) (Li et al., 2022b). ViT, in particular, revolutionizes the comprehension capabilities of the Context-Aware Visual Grounding (CAVG) model with its unique ability to consider the entire visual scene as a unified entity. Unlike conventional object-centric methods reliant on bounding boxes, ViT offers a comprehensive perspective of the visual scene. As depicted in Fig. 4, ViT captures essential contextual details—such as lane markings, pedestrian paths, traffic signs, and other nuances—that are often missed by rigid bounding box constraints. This expanded scope enables our model to assimilate a wider range of visual information and contextual cues, ensuring a thorough semantic interpretation.

In concert with ViT, BLIP acts as a bridge between textual and visual information, enabling CAVG to extract cross-modal features and make meaningful connections between the words in the command and the semantic elements

present in the image. This fusion of linguistic context with rich visual cues facilitates a more holistic and nuanced understanding of the scene, enhancing our model's ability to accurately ground textual commands in the visual scenes. The Context Encoder ingests the context vector $\boldsymbol{O}_{context}$ from text vector $\boldsymbol{O}_{text}$ and the input image I, which can be represented as follows:

$$\boldsymbol{O}_{context} = \phi_{BLIP}\left(\phi_{ViT}(I) \| \boldsymbol{O}_{text}\right) \tag{4}$$

where ϕ_{BLIP}, and ϕ_{ViT} are the BLIP and ViT frameworks, respectively. In addition, $\|$ denotes matrix concatenation and the context vector $\boldsymbol{O}_{context}$ not only augments the model with additional visual cues but also broadens its understanding of the scene. This, in turn, enhances the model's ability to make more accurate predictions.

3.6. Cross-Modal Encoder

This encoder synergizes the output of the previous encoders through a multi-head cross-modal attention mechanism. Unlike conventional methods like MaskFormer (Cheng et al., 2021), DETR (Carion et al., 2020), and MDETR (Kamath et al., 2021), which are computationally expensive, our approach uses a "hindsight fusion" strategy. This enables more efficient processing by harmoniously merging visual and text vectors before they are inputted into the Multimodal Decoder. As depicted in Fig. 5, the encoder projects the text vector $\boldsymbol{O}_{text}$ and the emotion vector $\boldsymbol{O}_{emo}$ into

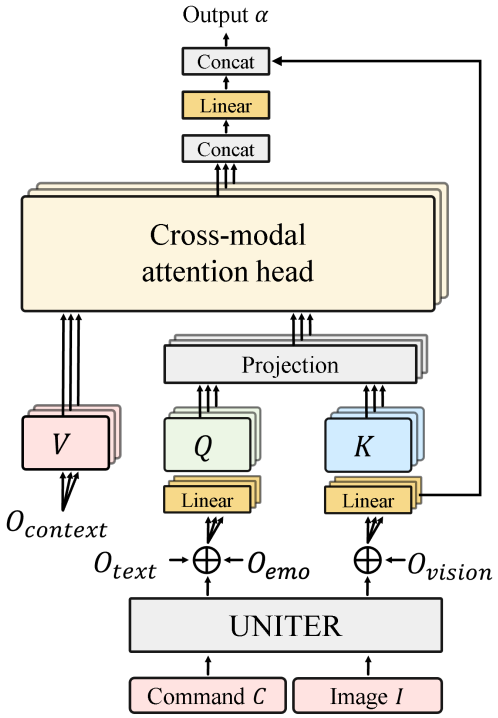


Figure 5: Overall Framework of Multi-head Cross-modal Self-attention Mechanism.

query representation Q using Multilayer Perceptrons (MLP). Similarly, the vision $\boldsymbol{O}_{vision}$ and context $\boldsymbol{O}_{context}$ vectors are converted into the key K and standardized value V representations, respectively. To further enhance the text and vision vectors, we append extra tokens L_q and L_k to the end of the input text vector $\boldsymbol{O}_{text}$ and vision vector $\boldsymbol{O}_{vision}$, respectively. Then, these representations are utilized to compute a comprehensive fused vector α, as outlined in the equations. In addition, it is worth noting that our cross-modal attention mechanism draws inspiration from the original Transformer (Vaswani et al., 2017) architecture and can also be extended to incorporate multiple attention heads, with a residual connection applied after these attention heads (Hao et al., 2019). Formally,

$$\begin{cases} Q_i = W^Q(\phi_{Linear}(\boldsymbol{O}_{vision} + L_q)) \\ K_i = W^K(\phi_{Linear}((\boldsymbol{O}_{emo} \| \boldsymbol{O}_{text}) + L_k)) \\ V_i = W^V \boldsymbol{O}_{context} \end{cases} \tag{5}$$

where W^Q, W^K, W^V denote the learnable weights, and ϕ_{Linear} represents the fully connected layer. These tokens L_q and L_k are generated as position embeddings from the given command C and the raw image I by leveraging the single-stream architecture UNITER (Universal Image-TExt Representation) (Chen et al., 2020), which can be expressed as follows:

$$\begin{cases} L_q = \phi_{UNITER}\ (C) \\ L_k = \phi_{UNITER}\ (I) \end{cases} \tag{6}$$

where ϕ_{UNITER} denotes the UNITER architecture. Moreover, the ith self-attention head $head_i$ and the output of the multi-head cross-modal self-attention mechanism α can be defined mathematically as:

$$\begin{aligned} \alpha &= \sum_{i=1}^{h} head_i + \phi_{\text{Linear}}(\boldsymbol{O}_{vision} + L_q) \\ &= \sum_{i=1}^{h} \phi_{softmax}\left(\frac{Q_i(K_i)^\top}{\sqrt{d_k}}\right) V_i + \phi_{Linear}(\boldsymbol{O}_{vision} + L_q) \end{aligned} \tag{7}$$

where h is the total number of attention heads, $\phi_{softmax}(\cdot)$ is the softmax activation function, and d_k represents the dimensionality of the projected key vectors. Furthermore, the output of this attention mechanism α is then embedded by the UNITER to produce the fused cross-modal vector $\bar{\alpha}$, which can be given as follows:

$$\bar{\alpha} = \phi_{UNITER}\ (\alpha) \tag{8}$$

Our methodology prominently features the UNITER architecture, renowned for its exceptional capacity to align and integrate visual and textual data into a unified representation. Built upon the transformer model, UNITER processes images and text in parallel, transforming images into patch embeddings analogous to text tokens. This method facilitates a seamless integration of linguistic and visual modalities, a crucial aspect for interpreting complex multimodal data in our study. Pretrained on a broad spectrum of V&L tasks, UNITER can discern a wide range of visual and textual patterns. This extensive pre-training is especially beneficial for our application, where interpreting the intricate interplay between natural language commands and corresponding visual cues is critical. UNITER's ability to create a joint representation of text and images is key to effectively grounding language commands within the specific visual contexts encountered by AVs.

3.7. Multimodal Decoder

The final stage of our architecture is the Multimodal Decoder, which plays a pivotal role in synthesizing and interpreting the multimodal data. This decoder integrates the high-level vector L_q, embedded by the UNITER, with the fused cross-modal vector $\bar{\alpha}$. Its primary function is to discern and rank the top-k RoIs in the image, thereby identifying the most relevant target region for the given command. Comprising m layers, the Multimodal Decoder is uniquely structured. Each layer is an amalgamation of a Transformer Decoder (Vaswani et al., 2017) and a Region-Specific Dynamic (RSD) Layer Attention Mechanism. The RSD mechanism, inspired by Chan et al. (Chan et al., 2022), represents an advanced evolution in attention mechanisms, designed specifically to address the multimodal complexity of our task.

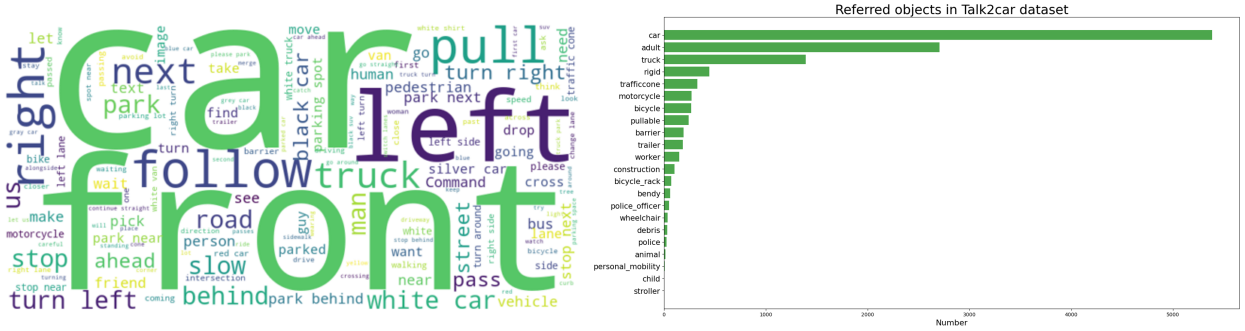
Within each layer, the Transformer Decoder generates a high-dimensional hidden state from its input. This state is then processed by the RSD Layer, a key component that enhances the model's interpretative capabilities. The RSD mechanism dynamically assigns attention weights to the representation vectors of a region across all encoder layers, rather than relying solely on the top layer representation. This approach allows for a more nuanced aggregation of features, ranging from concrete to abstract, across different layers of the encoder. For each region proposal v_i, the RSD Layer evaluates its representation $h_l^{v_i}$ at each encoder layer l, computing relevance scores α_{il}. These scores are used to aggregate the representation vectors across layers, enabling the model to consider a wide array of features for each region and determine its relevance to the command. The hidden states from all layers, each weighted by their computed relevance, are fused into a singular representation. This combined output is then passed through a Multilayer Perceptron (MLP) followed by a softmax activation function, resulting in multiple hypothesis outputs with associated credibility scores. The bounding box with the highest credibility score is selected as the final output. To further enhance feature retention and extraction, skip connections akin to those in ResNet (He et al., 2016) are employed. These connections

feed the high-level vector L_q into every layer of the Multimodal Decoder. This design allows the model to eliminate redundant features while preserving those essential for accurate interpretation and decision-making, thereby improving the accuracy and reliability of the model's output.

4. Experiments

4.1. Dataset

We assess CAVG using the Talk2Car (Deruyttere et al., 2019) dataset, a derivative of the NuScenes (Caesar et al., 2020) dataset, featuring 11,959 natural language commands across 9,217 images of urban environments from Singapore and Boston. The data set is particularly challenging and diverse, captured under various conditions including different times of the day and weather scenarios. It poses intricate tasks like object disambiguation and understanding elaborate text queries, with commands averaging 11 words in length, fitting CAVG's data requirements. Examples include complex instructions like "This car just passed a crosswalk, slow down to check for pedestrians!". We utilize it as a benchmark to assess CAVG's accuracy in this study.



(a) Word Cloud in Talk2car Dataset

(b) Number of Referred Objects

Figure 6: Statistics of Talk2car Datasets on (a) Word Cloud and (b) Distribution of Referred Object Number.

4.2. Evaluation Metric

In line with the C4AV challenge, we use the $IoU_{0.5}$ score as the evaluation metric. It is alternatively referred to as the AP_{50} score, which is defined as the percentage of Intersection over Union (IoU) score values that exceed the 0.5 thresholds with the ground truth bounding boxes. This evaluation metric quantifies is widely used in challenges and benchmarks such as the PASCAL (Everingham and Winn, 2012), VOC (Everingham et al., 2010), and COCO (Lin et al., 2014) datasets, provides a rigorous measure of the accuracy of our predictions, and is consistent with established practices in computer vision and object recognition tasks. Let A and B denote the predicted bounding box and the ground truth bounding box, respectively. In this study, each bounding box is represented by the coordinates of the positions of its upper-left (x_1, y_1) and lower-right (x_2, y_2) corners. The following equation elaborates on the computation of IoU between the predicted bounding box A and the ground truth bounding box B:

$$\text{IoU}(A, B) = \frac{\text{Area of Overlap}\,(A, B)}{\text{Area of Union}\,(A, B)} \tag{9}$$

4.3. Training and Implementation Details

Our model is trained to converge on an NVIDIA A100 40GB GPU for approximately an hour. It is trained for 6 epochs using the AdamW optimizer, with a batch size of 16 and an initial learning rate set at 10^{-4}. In addition, the loss function for CAVG is the binary cross-entropy (BCE) loss, with scheduling regulated by the Cosine Annealing Warm Restarts (Loshchilov and Hutter, 2016) strategy. Further specifics regarding the implementation and essential parameter settings of CAVG are provided as follows:

Text Encoder. In the text encoder, we use a pre-trained BERT model to generate text embeddings, enforcing a maximum sentence length of 60 tokens. We configure the BERT encoder with 16 hidden layers and an embedding

layer with a vocabulary size of 30,524. The epsilon value for the layer normalization is set to 1e-12. The hidden size within the text encoder is set to $d = 768$, resulting in a text vector labeled $\boldsymbol{O}_{text}$ with a dimension of d.

Emotion Encoder. In this encoder, GPT-4 is used as an emotion classifier to distinguish emotion categories and generate an embedding $\boldsymbol{O}_{emo}$ with a dimension of 768, ensuring semantic congruence with $\boldsymbol{O}_{text}$. We then concatenate $\boldsymbol{O}_{emo}$ and $\boldsymbol{O}_{text}$, projecting them to a dimension of d.

Vision Encoder. We resize the input image from 1600×900 to 800×450 in the Vision Encoder. Then, Fast R-CNN and ResNet-101 are used as the backbone to locate the bounding boxes of all the objects, and the region proposal threshold is set to 36. Within each region proposal, image information is extracted from the image feature $\boldsymbol{O}_{vision}$ with the dimension of 36×1024.

Context Encoder. The Context Encoder utilizes the complete raw image I in conjunction with the text vector $\boldsymbol{O}_{text}$. Within this framework, we choose an encoder between Vision Transformer-Base and Vision Transformer-Large for the vision embedding layer. Specifically, the ViT-B is configured with a depth of 12, a vision width of 768, a patch size of 16, a ratio of MLP hidden dimension to embedding dimension of 4, and a multi-head attention mechanism with 12 heads. The resulting encoder output is the context embedding $\boldsymbol{O}_{context}$ with a dimension of d.

Cross-Modal Encoder. The vectors of text $\boldsymbol{O}_{text}$, emotion $\boldsymbol{O}_{emotion}$, vision $\boldsymbol{O}_{vision}$, and context $\boldsymbol{O}_{context}$ vectors are fed into this encoder to be converted by the attention mechanism into representations of key, query and values, each with a dimension of 1024. The encoder is configured with 16 heads to optimize the attention mechanism. The UNITER then embeds both α and the text feature, using a hidden size of 1024 and producing an output of dimension d.

Multimodal Decoder. The Multimodal Decoder is designed with 12 attention layers. Within each layer, we utilize a multi-head attention mechanism configured with 12 heads, and both the vision feature and the text image have a hidden size set to d. Additionally, every attention layer incorporates a linear forward progression followed by batch normalization and dropout layers.

4.4. Experiment Setup

Our experimental protocol aligns with the Commands For Autonomous Vehicles (C4AV) Challenge protocol. We split the Talk2car dataset into 8,349 training, 1,163 validation, and 2,447 test samples, which constitute 69.8%, 9.7%, and 20.5% of the total dataset. We refer to the complete testset as the *full* testset. With this configuration in mind, we compute the distribution of the word counts for the natural language command on the Talk2Car dataset. The statistical results, shown in Fig. 7, indicate that the training, validation, and testsets exhibit similar data distribution profiles for input commands.

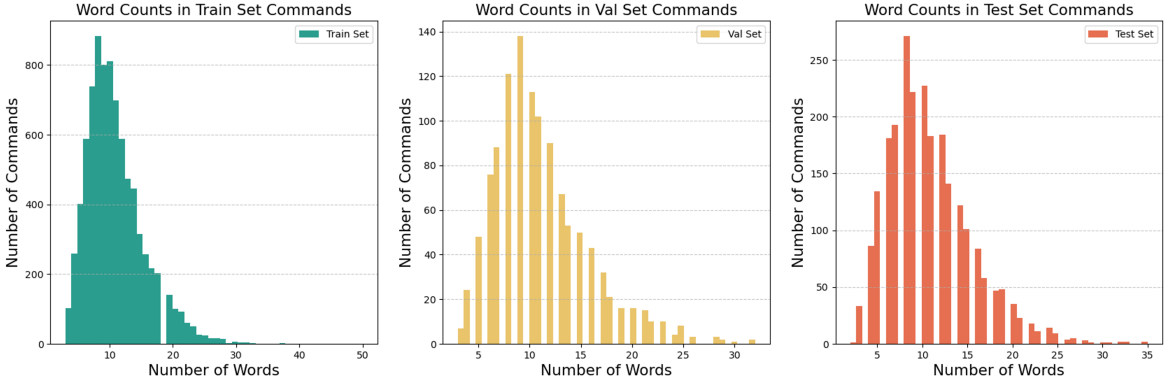


Figure 7: Word Count Distribution of Commands in the Training, Validation, and Test Set.

To rigorously evaluate the effectiveness of our model in real-world scenarios, we meticulously partitioned the validation set based on the complexity of linguistic commands and the inherent challenges of visual environments. It is worth noting that longer commands may introduce irrelevant details or be more complex for AVs to understand. In addition, specific visual scenarios - such as low-light night scenes, congested urban environments with complex object interactions, ambiguous command prompts, and rainy days where visibility is compromised - invariably complicate prediction. To this end, we prudently select more than 23 words in length and categorize the aforementioned

Table 1
Performance Comparison of Our Model with Various SOTA Methods on the Talk2Car Dataset using $IoU_{0.5}$ Metric. Models are categorized into three types: one-stage, two-stage, and others, and evaluated based on architectural backbones: *Vision* for visual feature extraction, *Language* for semantic information extraction, and *Global* for holistic data assimilation. Additional components evaluated include emotion classification (EmoClf.), global image feature extraction (Global Img Repr.), linguistic augmentation (NLP Augm.), and visual augmentation (Vis Augm.). 'Yes' indicates the use of a feature, 'No' its absence, and '-' represents undisclosed data. This categorization elucidates the underlying components and strategies influencing each model's performance. **Bold** and underlined values represent the best and second-best performance, respectively. For comprehensive results, visit the Official Leaderboard.

Model	$IoU_{0.5}$	Backbones			Methods		Augmentation	
		Vision	Language	Global	EmoClf.	Global Img Repr.	NLP Augm.	Vis Augm.
One-stage:								
MAC	50.5	ResNet-101	LSTM	ResNet+CNN	No	Yes	No	Yes
AttnGrounder	63.3	ResNet-53	LSTM	Darknet-53 (Redmon and Farhadi, 2018)	No	Yes	No	Yes
Two-stage:								
OSM	35.3	ResNet-18	GRU	ResNet-18	No	Yes	-	-
SCRC	43.8	ResNet-18	GRU	ResNet-18	No	Yes	-	-
Bi-Dir.retr	44.1	ResNet-18	LSTM	No	No	No	No	No
MSRR (Top-16)	59.9	ResNet-101	LSTM	No	No	No	No	No
MSRR	60.1	ResNet-101	LSTM	No	No	No	No	No
ASSMR (4th)	66.4	ResNet-18	GRU	ResNet-101	No	Optional	Yes	Yes
CMSVG (3rd)	68.6	EfficientNet (Tan and Le, 2019)	Sent.-Transf (Reimers and Gurevych, 2019)	No	No	No	No	No
CMRT (2nd)	69.1	ResNet-152	Transformer (Vaswani et al., 2017)	ResNet-152	No	Yes	-	-
Sentence-BET	70.1	ResNet-152	Sent.-Transf (Reimers and Gurevych, 2019)	ResNet-152	No	Yes	-	-
Stacked VL-BERT (1st)	71.0	ResNet-101	VL-BERT	ResNet-101	No	Yes	-	-
Others:								
C4AV-Base	44.3	-	-	-	No	-	No	-
A-AAR	45.1	VGG-16(Simonyan and Zisserman, 2014)	Hier.-LSTM (Lu et al., 2016)	-	No	Yes	-	-
MaskObjs (Top-16)	54.3	-	-	-	No	-	-	-
VL-BERT (base)	68.3	-	-	-	No	-	-	-
SOTA (Top 1-6)	66.4±5.6							
SOTA (Top 7-12)	50.1±6.8							
CAVG (50%)	70.3	CenterNet	BERT/BLIP	UNITER/BLIP	Yes	Yes	Yes	Yes
CAVG (75%)	72.1	CenterNet	BERT/BLIP	UNITER/BLIP	Yes	Yes	Yes	Yes
CAVG	**74.6**	CenterNet	BERT/BLIP	UNITER/BLIP	Yes	Yes	Yes	Yes

challenging scenarios into specialized subsets, designated as the *Long-text* and *Corner-case* testsets, respectively. This categorization allows for more focused evaluations.

In addition, the *Corner-case* test set consists of three main subsets: the scenarios with restricted vision, the scenarios with multi-agent interactions, and the scenarios with ambiguous command prompts, containing 65, 75, and 85 samples, respectively. Figure 8 displays the example scenes for each subset; it is pertinent to mention that a single sample can simultaneously span multiple subsets, provided it is consistent with the definition of multiple scenarios. Correspondingly, we treat the C4AV test set as a "normal" subset for comparative analysis. Furthermore, for the *Long-text* test set, we employed a data augmentation strategy to increase the richness of the dataset without deviating from the original semantic intent. Using LLM GPT-3.5, we extended certain commands, resulting in commands ranging from 23 to 50 words. This allows us to further evaluate the model's ability to process extended linguistic input, ensuring a comprehensive assessment of our model's adaptability and robustness.



Command: Pull over at this speedbump so I can ask those people where the restaurant is.

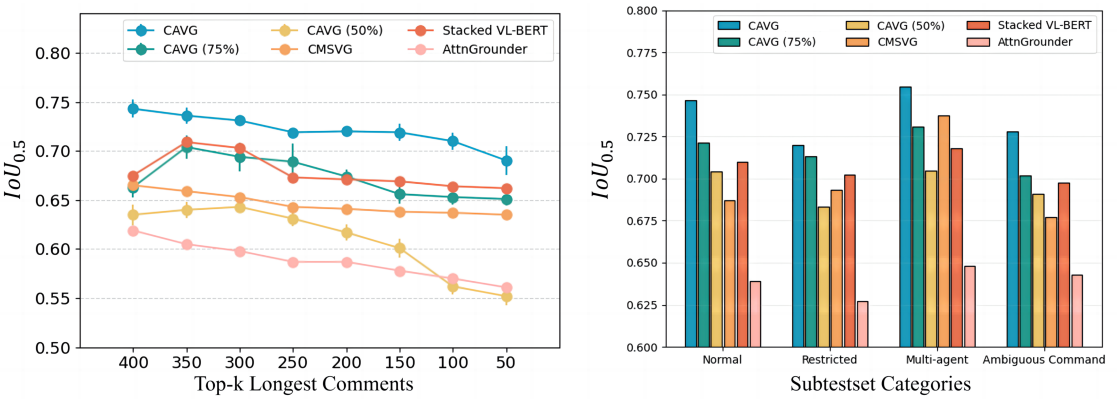
Command: Pull up next to the ford SUV in the left lane.

Command: My friend is the guy standing closest to the curb, next to that car in front of us. Pull over so he can get in.

Figure 8: Example for the Subsets of the *Corner-case* Test Set.

4.4.1. Performance Benchmarking Against SOTA Models on full Testset

In our research, we rigorously evaluated CAVG's performance against a comprehensive array of state-of-the-art (SOTA) methods. These include well-known benchmarks such as STACK (Hu et al., 2016), OSC (Vandenhende



(a) Long-text Testset (b) Corner-case Testset

Figure 9: Evaluation Results in *Long-text* and *Corner-case* Testsets.

et al., 2020), SCRC (Vandenhende et al., 2020), C4AV-Base (Hudson and Manning, 2018), A-ARR (Deng et al., 2018), MAC (Hudson and Manning, 2018), MaskObjs (Top-16) (Deruyttere et al., 2020a), MSRR (Deruyttere et al., 2020a), AttnGrounder (Mittal, 2020), VL-BERT (Base) (Vandenhende et al., 2020), and Sentence-BET (Grujicic et al., 2022). Additionally, we included the top four contenders from the prestigious C4AV Challenge (Deruyttere et al., 2019)—namely, Stacked VL-BERT (1st) (Dai et al., 2020), CMRT (2nd) (Luo et al., 2020), CMSVG (3rd) (Rufus et al., 2020), and ASSMR (4th) (Ou and Zhang, 2020)—to ensure a holistic assessment. The complete performance metrics for this comparative analysis are available in Table 1.

Remarkably, within the *full* testset, our model did not just hold its own—it dominated across the board. Not only did it surpass all the aforementioned SOTA methods, but it also eclipsed the performances of all models from the C4AV Challenge leaderboard. In terms of the $IoU_{0.5}$ metric, the superiority of CAVG is clear: it outperformed Stacked VL-BERT (1st) by a margin of 4.8%, outperformed CMRT (2nd) by 7.4%, eclipsed CMSVG (3rd) by 7.9%, and dramatically outperformed ASSMR (4th) by an impressive 11.0%. This compelling empirical evidence strongly attests to the exceptional efficacy of our proposed model in the task at hand.

4.4.2. Performance Evaluation of Training on Reduced Datasets

In order to showcase the scalability and efficiency of CAVG, we conducted an experiment in which the model was trained on different subsets of the training set, specifically 50% and 75%. As illustrated in Tables 1, under the labels CAVG (50%) and CAVG (75%), CAVG exhibits commendable performance. When conditioned on the 75% subset of the training data, CAVG consistently outperforms all other baseline models, outperforming C4AV SOTA baselines Stacked VL-BERT (1st) and CMRT (2nd) by 5.07% and 7.96%, respectively. Impressively, an interesting finding is that when the model is trained on only 50% of the training data, CAVG displays an admirable performance by achieving significantly lower $IoU_{0.5}$ values compared to most of the baselines, and outperforms almost all of the top 1-6 SOTA baselines. These results underscore the potential of CAVG to provide competitive performance even in the absence of extensive training data. This property is of paramount importance, as it suggests potential avenues for economizing data acquisition and annotation efforts, particularly relevant for the autonomous vehicle industry, where data acquisition can be both time-consuming and costly.

4.4.3. Performance Evaluation on Long-text and Corner-case Testsets

For a holistic insight into the robustness and adaptability of CAVG, we performed further evaluations on the *Long-text* and *Corner-case* testsets, as detailed in Table 2 and Table 3, respectively. To our knowledge, few methods in this area provide the code for these proposed models. Therefore, our comparison focuses on competing open-source models. When confronted with the challenges of the *Long-text* testset, as shown in Fig. 9 (a), CAVG consistently surpasses all other baselines, showing an improvement of 4.0%-10.1% over the SOTA baselines. Notably, even our model variant trained on only 75% of the dataset, CAVG (75%), shows commendable performance, rivaling if not surpassing the current SOTA model, Stacked VL-BERT (1st). Furthermore, the iteration of CAVG trained with only half the dataset, CAVG (50%), shows impressive performance, outperforming the majority of the established models. Interestingly,

Table 2
Performance Comparison of Our Model with Open-source SOTA Methods on the *Long-text* Testset using $IoU_{0.5}$ Metric. **Bold** and underlined values represent the best and second-best performance, respectively.

Model	Top-k Longest Comments							
	400	350	300	250	200	150	100	50
AttnGrounder	61.9	60.5	59.8	58.7	58.6	57.8	57.0	56.1
CMSVG	66.5	65.9	65.3	64.3	64.1	63.8	63.7	63.5
Stacked VL-BERT (1st)	67.5	70.9	70.3	67.3	67.1	66.9	66.4	66.2
CAVG (50%)	63.5	64.0	64.3	63.1	61.7	60.1	56.2	55.2
CAVG (75%)	66.3	70.4	69.4	68.9	67.4	65.6	65.3	65.1
CAVG (100%)	**74.1**	**73.6**	**73.1**	**71.9**	**72.2**	**71.9**	**71.0**	**69.1**

Table 3
Performance Comparison of Our Model with Open-source SOTA Methods on the *Corner-case* Testset using $IoU_{0.5}$ Metric. **Bold** and underlined values represent the best and second-best performance, respectively.

Model	Subtestset Categories			
	Normal	Restricted	Multi-agent	Ambiguous Command
AttnGrounder	63.9	62.7	64.8	64.3
CMSVG	68.7	69.3	73.8	67.8
Stacked VL-BERT (1st)	71.1	70.2	71.8	69.8
CAVG (50%)	70.4	68.8	70.4	69.1
CAVG (75%)	72.1	71.3	73.1	70.2
CAVG (100%)	**74.6**	**72.1**	**75.2**	**72.8**

our models - CAVG and CAVG (75%) - excel in multi-agent scenarios compared to normal ones when provided with sufficient training data. These results highlight their superior ability to understand and interpret commands of considerable length and complexity.

Shifting our gaze to the *Corner-case* testset in Fig. 9 (b), the strengths of CAVG become even more apparent in the challenging scenes. It exceeds current SOTA baselines, with accuracy gains ranging from 1.9%-4.9% on the Talk2car dataset. Moreover, even the CAVG trained on only 75% of the dataset outperforms other SOTA baselines, while its counterpart, the CAVG trained on only 50% of the dataset, still manifests commendable performance. This demonstrates the inherent robustness of CAVG, even under challenging conditions, and underscores its promise for use in real-world AV scenarios.

4.4.4. Comparative Analysis of Model Inference Speed

We evaluate the inference speed of CAVG models for different architectures in the Talk2car dataset. Specifically, we test the inference speed of all models on a single NVIDIA GeForce RTX 3090 24GB GPU for a fair comparison. Table 4 presents a comparative analysis of various models against the CAVG (baseline), which employs the standard R-CNN (Girshick et al., 2014) as a replacement for the Vision Encoder, omitting the BLIP and Context Encoder. These models are defined as follows, with each compared to the full proposed model CAVG:

- CAVG: The full model proposed in this paper.
- CAVG (ResNet-152): Uses ResNet-152 in the Vision Encoder for visual feature extraction.
- CAVG (ResNet-50): Includes ResNet-50 in the Vision Encoder for vision vector extraction.
- CAVG (ResNet-18): Includes ResNet-18 in the Vision Encoder for vision vector extraction.
- CAVG (BLIP): Applies BLIP only in the Text and Context Encoders for vision and context vector extraction.
- CAVG (Small): A scaled-down model featuring a reduced number of attention heads (4 heads) in the Cross-Modal Encoder with 256d hidden states, and fewer attention layers (4 layers) in the RSD.

Table 4
Inference Time of CAVG Models for Different Architectures and their Predicted Performance in the Talk2car dataset. CAVG (baseline) is selected as the baseline for comparative analysis. These models are evaluated based on architectural backbones: *Vision* for visual feature extraction, *Language* for semantic information extraction, and *Global* for holistic data assimilation. **Bold** and underlined values represent the best and second-best performance, respectively.

Model	Accuracy		Speed			Backbones		
	$IoU_{0.5}$	Accuracy-up (%)	Sample Size	Inference Time (s)	Speed-up (%)	Vision	Language	Global
CAVG (baseline)	61.4	-	64	2.43212	-	R-CNN	BERT	-
CAVG (ResNet-152)	74.1	20.68%	64	1.94372	20.1%	ResNet-152	BERT/BLIP	UNITER/BLIP
CAVG (ResNet-50)	72.2	17.59%	64	1.74321	28.3%	ResNet-50	BERT/BLIP	UNITER/BLIP
CAVG (ResNet-18)	68.8	12.05%	64	<u>1.66528</u>	<u>31.5%</u>	ResNet-18	BERT/BLIP	UNITER/BLIP
CAVG (BLIP)	71.4	16.28%	64	1.77284	27.1%	ResNet-101	BLIP	BLIP
CAVG (Small)	<u>72.5</u>	18.08%	64	**0.90134**	**62.9%**	ResNet-18	BERT/BLIP	UNITER/BLIP
CAVG	**74.6**	**21.50%**	64	1.90446	21.7%	ResNet-101	BERT/BLIP	UNITER/BLIP

We obtain the average inference time of CAVG (baseline) is about 0.037s per sample. Meanwhile, using the Vision Encoder can speed up the inference speed, showing at least improvements of 20.1% compared to the baseline. This result demonstrates that incorporating CenterNet into the Vision Encoder significantly increases processing speed, while achieving better performance in terms of prediction accuracy, underscoring its effectiveness in visual grounding contexts for AVs where accuracy, real-time responses are critical. Despite utilizing smaller ResNet variants in the Vision Encoder, the CAVG (ResNet-18) and CAVG (ResNet-50) models maintain strong predictive capabilities, outshining many SOTA models in the *full* test set. They also significantly boost inference speeds by 31. 5% and 28. 3%, respectively, compared to CAVG (baseline). Intuitively, the inference time increases with the larger size of the ResNet used in the Vision Encoder. However, the CAVG (ResNet-152), despite its larger size, still provides faster inference capabilities than the CAVG (baseline). It is noted that the CAVG (Small) achieves an optimal balance between computational efficiency and prediction accuracy. This model is 62.9% faster in inference than the CAVG (baseline), while also improving prediction performance by 2.1% compared to Stacked VL-BERT (1st). This highlights the potential of the proposed model to achieve high prediction accuracy with significantly reduced computational requirements.

4.5. Case Study for RSD Layer Attention Weights

We visualize the distribution of layer attention weights for 13 layers in a multimodal decoder to further showcase the effectiveness of our proposed RSD layer attention mechanism. We classify the input regions into two distinct groups based on their IoU alignment with the ground-truth region:

- $IoU > 0$: encompassing all regions with an IoU exceeding 0, indicating an overlap with the ground-truth region.
- $IoU = 0$: Constituting regions exhibiting no overlap, having an IoU of precisely 0.

As shown in Fig. 10, the higher decoder layers (especially layers 7 to 10) are endowed with a larger fraction of attention weights. This observation suggests that vectors have a greater influence on these higher layers, possibly due to increased cross-modal interactions. Interestingly, contrary to our initial expectations, the top layer does not dominate the attention weights. On the contrary, its output appears to be less central. This is a marked departure from traditional techniques that rely predominantly on the topmost layer representation to predict the best-aligned region, potentially sidelining the nuanced cross-modal features inherent in other layers. In light of this, we introduce the RSD layer mechanism, which is designed to dynamically allocate attention weights across all decoder layers, anchored by their hidden states. Subsequently, the model integrates the outputs of different layers and harmonizes them on the basis of these attention weights. This innovative approach equips our model with the ability to capture and exploit richer cross-modal semantic features, which are essential to command and scene understanding.

4.6. Ablation Study

4.6.1. Ablation Study for Core Module

We conducted a rigorous ablation study to determine the effect of core components in CAVG. As shown in Table 5, specifically, model H, inclusive of all components, consistently outperforms across the evaluation metric, illustrating the collective value of the components for optimal performance. Model A, a reduced version of Model H, replaces the

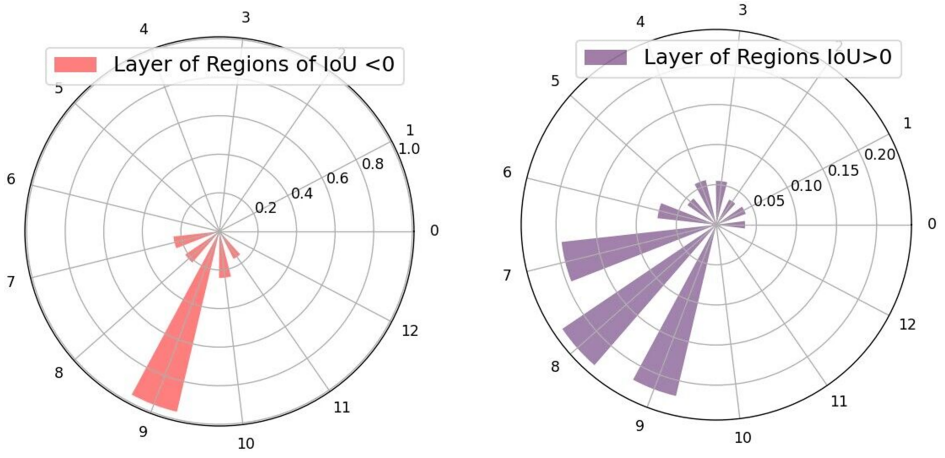


Figure 10: Attention Weights over all Decoder Layers for RSD Layer Attention. The 0-th layer indicates the embedding layer, while the 1-st denotes the first layer of RSD layer attention, and the 12-th represents the last layer.

Table 5
Component Constituents of Ablation Models.

Components	Ablation Models							
	A	B	C	D	E	F	G	H
Vision Encoder	✘	✔	✔	✔	✔	✔	✔	✔
Context Encoder	✔	✘	✔	✔	✔	✔	✘	✔
Text Encoder	✔	✔	✘	✔	✔	✔	✔	✔
Cross-Modal Encoder	✔	✔	✘	✘	✔	✔	✔	✔
UNITER-Embedding	✔	✔	✔	✔	✘	✔	✘	✔
Emotion Encoder	✔	✔	✔	✔	✘	✘	✘	✔
$IoU_{0.5}$	71.3	66.5	68.4	70.1	71.9	73.2	64.1	74.6

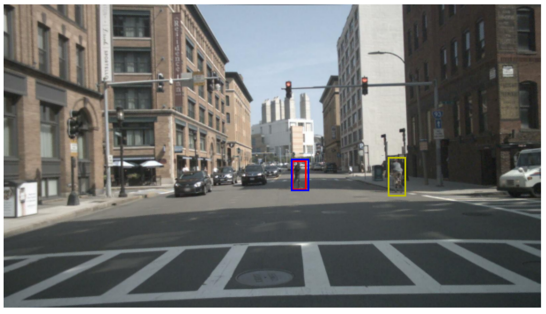
Vision Encoder with the classical R-CNN, leading to substantial decreases in $IoU_{0.5}$ score by 5.8%, highlighting the importance of visual features in prediction accuracy. This result also reveals the remarkable superiority of the Vision Encoder structure over the traditional R-CNN model. The Vision Encoder, with its single-stage, anchor-free design, streamlines the object detection process. It locates object centers directly on the feature map and uses regression to determine object dimensions. This approach not only eliminates the need for region proposal and anchor adjustment, simplifying the detection process but also provides a significant increase in detection accuracy. Model B, which excludes the Context and Emotion Encoder compared to the full model, shows significant metric reductions, this underscores the essential role of the additional visual cues in improving prediction accuracy. Furthermore, Model C, which omits the Text Encoder and Cross-Modal Encoder, exhibits performance drops of 6.6% in metric, suggesting the importance of processing instructions and multimodal information individually in this task. In addition, Model D, which lacks the visual Text Encoder, suffers notable performance degradation, emphasizing the improved efficacy of the proposed multi-head cross-modal attention mechanism. Accordingly, Model E and Model F omit the UNITER embedding process in the Multimodal Decoder and the Emotion Encoder, respectively, resulting in a decrease in the $IoU_{0.5}$ score as well. Finally, Model G, which omits the three core components of CAVG, suffers a noticeable decline in performance compared to Model H. This demonstrates the significant role that the synergistic integration of various modules plays in enhancing the accuracy of our model.

4.6.2. Case Study for the Context Encoder and Cross-modal Attention

To showcase the effectiveness of the proposed Context Encoder and the cross-modal attention mechanism and their interpretability. We visualize the prediction results made by CAVG with and without these components in Fig. 11. Specifically, we select several challenging examples where CAVG, without the Context Encoder and cross-modal attention mechanism, made inaccurate predictions. However, the inclusion of these components markedly improved



Command: Pull up alongside the curb. We need to pick up the pedestrian on the right that is closest to us.



Command: Watch for the pedestrian when you pass, I do not think they are supposed to be crossing there.

Figure 11: Visualization of Prediction Results for CAVG with and without the Integration of Context Encoder and the Cross-Modal Attention Mechanism. The blue box denotes the ground-truth region of interest. The yellow bounding box represents predictions made by CAVG without the use of a context encoder and multi-head cross-modal attention mechanisms. The red box shows predictions from CAVG when these advanced components are incorporated.

the predictive accuracy of CAVG, as evidenced by our thorough evaluation. For a more detailed understanding, Fig. 12 shows the output of the raw image after processing by the Context Encoder. This processing emphasizes key scene semantics in traffic scenarios, such as lane boundaries, road signs, and pedestrian paths. Moreover, the Context Encoder is able to capture more nuanced environmental cues, such as variations in lighting or indications of weather conditions. This extends the visual range of the model beyond the bounding boxes to a more holistic view of the environment. In addition, the cross-modal attention mechanism is instrumental in effectively fusing data from different modalities. Thus, the model goes beyond the visual input and harmonizes it with the given command.

4.6.3. Subtle Value Analysis of the Emotion Encoder

Our ablation study, specifically focusing on the Model F without the Emotion Encoder, provides a revealing perspective. At first glance, the empirical improvements attributed to the inclusion of this encoder may seem relatively modest when measured against the broader testset. However, a deeper dive reveals the critical role of this component. We argue that the power of the Emotion Encoder is not captured by broad metrics alone; its true potential emerges in specific, critical scenarios. For example, in a real-world driving context, consider a situation in which a visibly anxious passenger commands the AV to "drive fast"; while the literal interpretation focuses on speed, an understanding of the underlying emotional urgency might prompt the AV to prioritize a destination with fewer obstacles. Such human-friendly responses, attuned to human emotions, could significantly improve passenger comfort. Although these emotion-laden scenarios might constitute a smaller fraction of the entire dataset, their impact is disproportionately significant.

Importantly, the inclusion of the Emotion Encoder is strategic and lays the groundwork for our future efforts. Grounding linguistic commands in a visual context is only the first step. The next frontier, armed with this understanding of emotional context, is seamless trajectory planning that not only adheres to the explicit command but also resonates with the emotional state of the commander. In essence, by integrating the Emotion Encoder, we expect to evolve CAVG toward truly empathetic and anticipatory autonomous driving systems.

4.7. Intuition and Interpretability Analysis

CAVG is constructed with several key design motivations. First, a context encoder captures essential semantic relationships between bounding boxes by ingesting raw images and textual vectors, thereby enriching cross-modal understanding. Second, we present an innovative Emotion Encoder that analyzes the emotional information implicit in commander commands to more deeply capture human-AV interactions. Furthermore, we introduce a novel cross-modal attention mechanism to harmonize different types of input data. Finally, the RSD layer attention mechanism is employed to fine-tune the fusion of multi-modal vectors and hidden states, enhancing feature interpretability. These combined elements make CAVG both intuitive and adept at making robust predictions in complex scenarios.

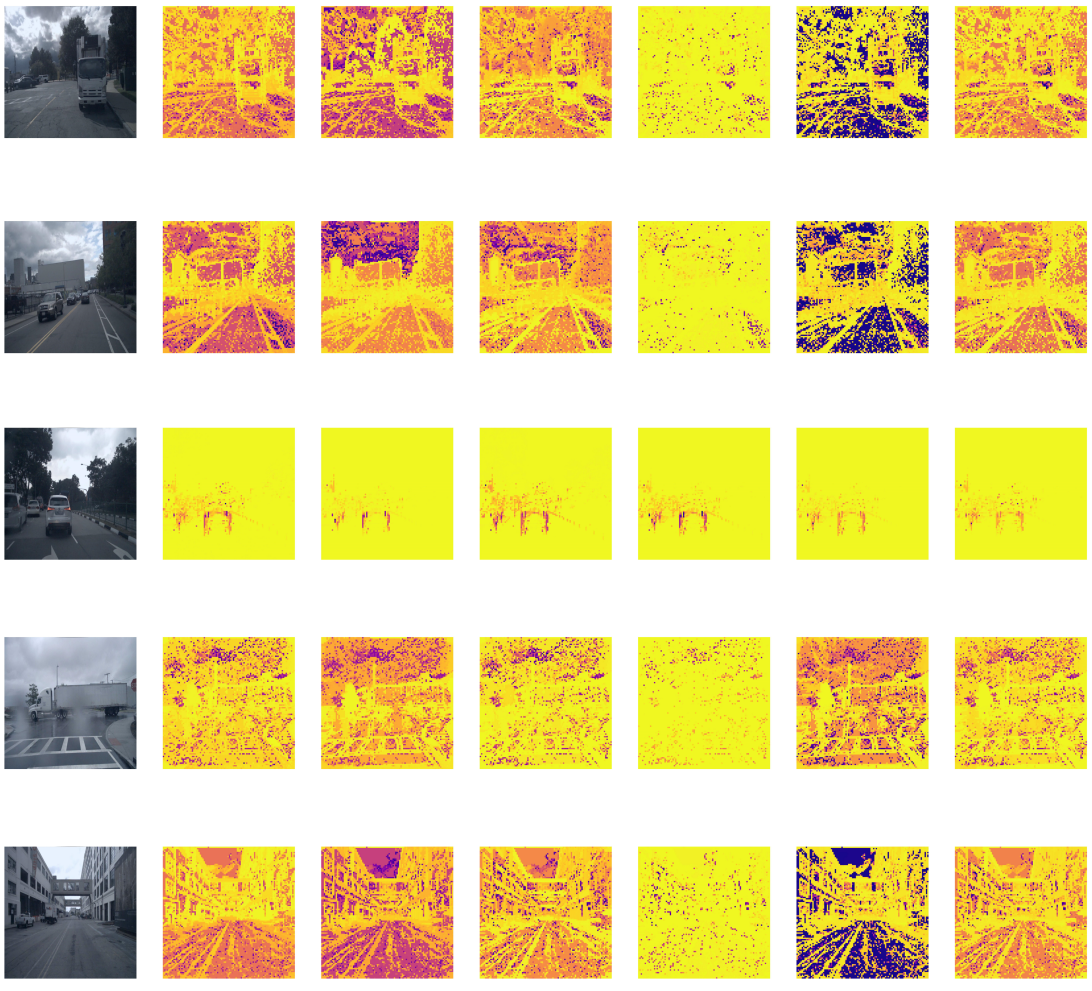


Figure 12: Raw Images and their Self-Attention Maps Processed by the Context Encoder. The raw images are visualized on the left, while the attention maps of these images in different layers within the encoder are displayed on the right.

Figure 13 vividly illustrates how our innovative emotion classification paradigm provides indispensable guidance to CAVG in certain contexts. Given a commander's command "Oh, right here! Park behind that white car.", the ambiguity inherent in such a directive becomes clear. Without integrating the commander's emotional context, the model tends to predict parking in front of the approaching white vehicle, resulting in a notable deviation from the ground truth. In contrast, our novel Emotion Encoder delves deep into the linguistic milieu of the command. Through nuanced linguistic analysis, such as recognizing the urgency conveyed by exclamatory phrases or specific keywords, it can identify underlying emotional undertones. In this case, the Emotion Encoder categorizes the emotion as **"urgent"**. This classification, when fused with the text vector produced by the Text Encoder, produces a richer emotion-sensitive embedding. This fusion communicates to the model the paramount importance and urgency of the commander's request. With this emotional insight, the model can better understand the commander's immediate needs. This enhanced understanding allows for the mitigation of potential misjudgments caused by ambiguous commands while taking into account human emotions for a more human-centered response.

In addition, in Fig. 14, we demonstrate the crucial role of cross-modal attention mechanisms in disambiguating natural language commands. Specifically, the attention to certain keywords in the command can greatly influence the model's output. For example, in the depicted scenario, the term **"just"** serves as a key keyword. Overlooking this keyword introduces ambiguity into the system, causing the model to consider multiple vehicles that meet the given criteria. Our multi-modal attention mechanism addresses this by correlating word vectors from the command with

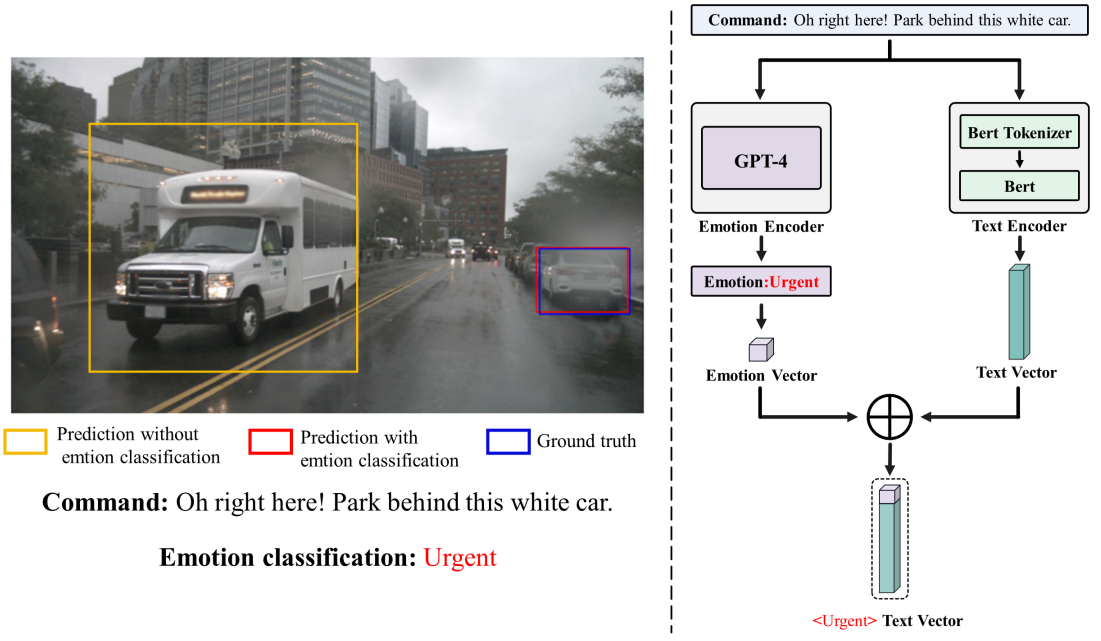


Figure 13: Illustration of the Emotion Encoder. Demonstrate how the Emotion Encoder indicates the model to identify a specific vehicle using an emotion vector embedded in a text vector.

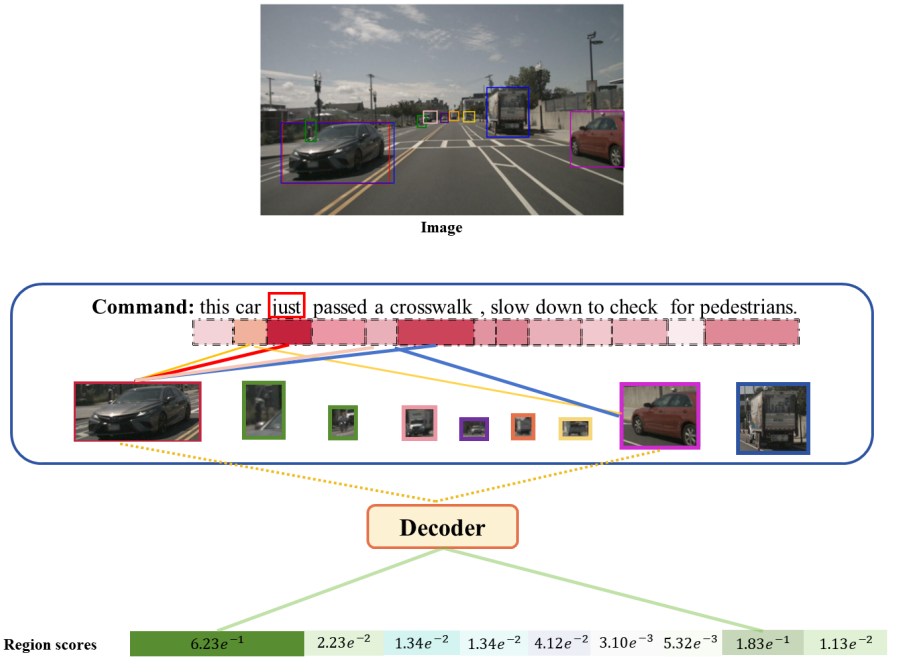


Figure 14: Illustration of the Cross-Modal Attention Mechanism. Demonstration of how keywords such as **"just"** guide the model in identifying specific vehicles, using word vectors and regional features.



Figure 15: Comparative Visualization of Model Performance on the Talk2Car Dataset. Ground truth bounding boxes are depicted in blue, while output bounding boxes of CAVG are highlighted in red. A natural language command associated with each visual scenario is also displayed below the image for context.

regional feature vectors in the image. This allows the model to allocate varying degrees of importance to different segments of the command. For example, only when the model incorporates global features like zebra crossings can it understand that a vehicle has **“just”** passed the crossing. Consequently, regions that are more semantically aligned with the keywords in the command receive higher attention scores. This enhanced attention mechanism captures not just the features within the bounding boxes but also broader contextual elements, such as lane markings and zebra crossings, allowing the model to make more informed and precise decisions.

4.8. Qualitative Results

To demonstrate the reasoning capabilities of our model, we select a set of challenging driving scenarios for more visual inspection. These selected scenarios exemplify the diverse and multifaceted situations that a vehicle may encounter. The visual demonstrations of our model’s performance on these instances, derived from the Talk2car dataset, are shown in Fig. 15 and Fig. 16. CAVG shows remarkable performance in predicting semantic regions from the graphical representation, demonstrating its robustness over a range of environmental conditions. These include challenging conditions such as low-light night scenes, urban environments with complex object interactions, ambiguous command scenes, and rainy days where visibility is compromised.

5. User Study

To assess user satisfaction with our model and SOTA models in real-world applications, we conducted a comprehensive user study using questionnaires. The purpose of this study is to gather user experience and feedback on various visual grounding models for AVs. By analyzing their responses, we sought to evaluate these models in detail and identify their strengths and areas for improvement. This user-centered approach reveals the real-world applicability and effectiveness of these models and contributes to the development of user-friendly AV systems.

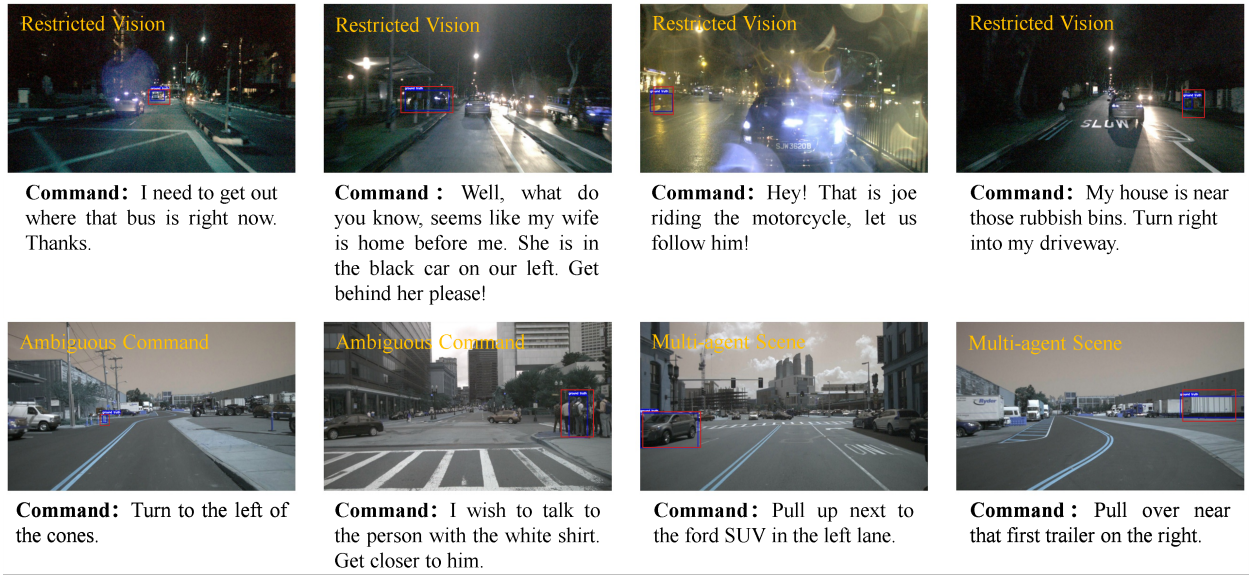


Figure 16: Comparative Visualization of Model Performance on Challenging Scenes. The challenging scenes include those with limited visibility, ambiguous commands, and scenes with multiple agents. Ground truth bounding boxes are depicted in blue, while output bounding boxes of CAVG are highlighted in red. A natural language command associated with each visual scenario is also displayed below the image for context.

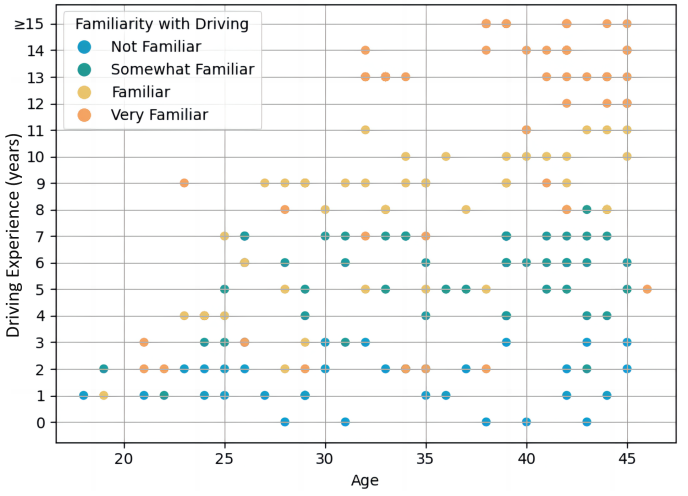
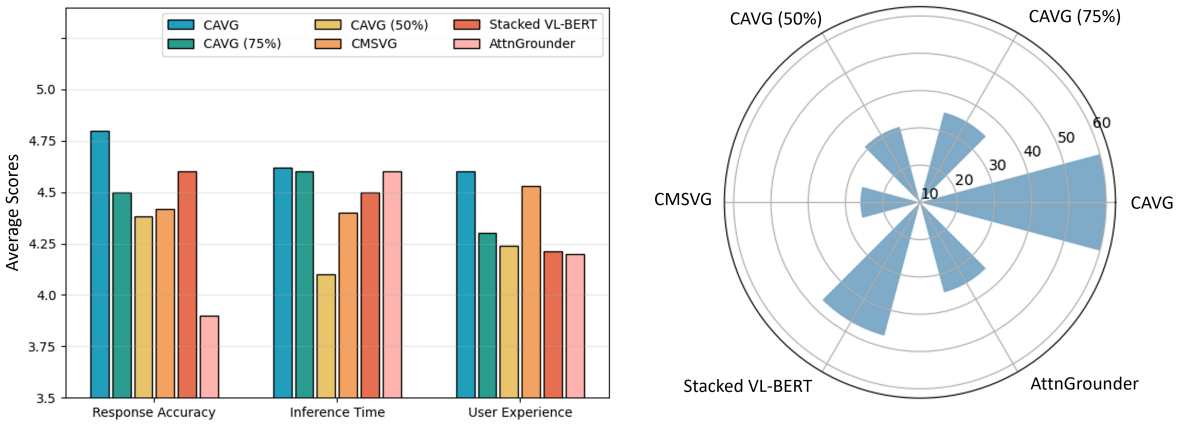


Figure 17: Age and Driving Experience Distribution of Survey Participants.

Recognizing that user satisfaction is closely related to the driving experience and skills of the participants, our survey included a diverse group of 174 participants, categorized by gender, age, driving experience, and educational background. As shown in Fig. 17, the group of participants is evenly distributed and includes users with varying levels of driving experience, including those with extensive experience (over ten years) and those with less experience (one to three years). These participants evaluate five different visual grounding models (CAVG, CAVG (75%), CAVG (50%), Stacked VL-BERT, CMSVG, and AttnGrounder) based on three key performance indicators (response accuracy, inference time, and user experience), and select the most preferred model from among them. Notably, to ensure fairness, the survey is anonymous; participants are unaware of the specific categories of the five models. Details of the survey can be found in the Appendix.



(a) (b)

Figure 18: Evaluation Results for the User Study. (a) Average scores of the models in response accuracy, inference time, and user experience. (b) Quantitative Distribution of favorite models selected by survey participants.

Figure 18 (a) shows a five-level grading system for the three key user experience metrics, with scores ranging from 0 to 5. CAVG scores highest in terms of response accuracy and user experience alignment, while also maintaining the faster response time. In addition, both the CAVG and CAVG (75%) provide an impressive user experience and receive relatively high user ratings compared to other models. Overall, CAVG outperforms the existing models by 5.2% to 10.6% in terms of comprehensive user ratings. Figure 18 (b) shows the participants' preferences for the five models. Among the preferred models, CAVG receives the most favor from the participants, indicating CAVG's practicality and significantly higher user satisfaction compared to other models. Crucially, feedback from the survey participants indicates that our future work should focus on further optimizing the structure of the model. This would involve balancing prediction accuracy with improved inference speed to respond more promptly to user commands and meet the real-time demands of visual grounding for autonomous driving, where immediate and accurate response is paramount for safety and user satisfaction.

6. Conclusion

In this research, we have developed a CAVG framework, meticulously crafted to enhance visual grounding in AVs. This framework is an integration of five core encoders—Text, Emotion, Vision, Context, and Cross-Modal—each contributing uniquely to the processing of multi-modal data, alongside a Multimodal Decoder. The Text Encoder, utilizing the BERT framework, plays a crucial role in extracting the linguistic nuances within commander commands. It ensures that the subtleties of language are accurately interpreted. Complementing this, our Emotion Encoder, leveraging the capabilities of GPT-4, is pivotal in understanding and classifying the emotional context of commands. This feature enables AVs to respond to human commands in a more empathetic and contextually aware manner. The Vision Encoder is a key component of our model, going beyond conventional object recognition. It employs frameworks like ViT and BLIP to provide a richer contextual understanding of traffic scenes. This capability is crucial for accurately interpreting and responding to grounding commands with a broader range of visual information. Furthermore, the Cross-Modal Encoder introduces a multi-head cross-modal attention mechanism. This mechanism adeptly combines textual and visual information, focusing attention on the most pertinent vectors. The Multimodal Decoder then employs RSD layer attention to fuse insights from each layer, leading to informed multimodal predictions. Tested on the Talk2Car dataset, CAVG demonstrates exceptional prediction accuracy, surpassing existing state-of-the-art methods. Notably, it maintains superior performance even with limited training data, ranging from 50% to 75% of the full dataset. Its robustness is further evident under challenging scenarios such as low-light night scenes, long-text command scenarios, rainy days, and dense urban streets.

In conclusion, CAVG not only sets a new benchmark in the AV domain but also significantly enhances the safety and human-centric aspect of the autonomous driving experience. Future research will focus on expanding CAVG's

capabilities, including integrating additional modalities like bird's-eye-view cues, lidar feedback, and HD maps. We also plan to investigate CAVG's integration into a comprehensive AV framework, particularly aligning with trajectory planning tasks.

Acknowledgement

This research is supported by the Science and Technology Development Fund of Macau SAR (File no. 0021/2022/ITP, 0081/2022/A2, SKL-IoTSC(UM)-2021-2023/ORP/GA08/2022, SKL-IoTSC(UM)-2024-2026/ORP/GA06/2023), and University of Macau (SRG2023-00037-IOTSC). We extend our gratitude to the Institute for Advanced Studies in Humanities and Social Sciences (IAS) for their valuable assistance. For any correspondence regarding this paper, please contact Dr. Zhenning Li at zhenningli@um.edu.mo or Dr. Chengzhong Xu at czxu@um.edu.mo.

CRediT authorship contribution statement

Haicheng Liao: Conceptualization, Methodology, Experiment, Writing. **Huanming Shen:** Methodology, Experiment, Writing. **Zhenning Li:** Conceptualization, Methodology, Writing. **Chengyue Wang:** Experiment. **Guofa Li:** Methodology. **Yiming Bie:** Methodology. **Chengzhong Xu:** Conceptualization, Funding, Review.

References

Bhattacharyya, A., Mauceri, C., Palmer, M., Heckman, C., 2022. Aligning images and text with semantic role labels for fine-grained cross-modal understanding, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, pp. 4944–4954.

Bonnefon, J.F., Shariff, A., Rahwan, I., 2016. The social dilemma of autonomous vehicles. Science 352, 1573–1576.

Bugliarello, E., Cotterell, R., Okazaki, N., Elliott, D., 2021. Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language berts. Transactions of the Association for Computational Linguistics 9, 978–994.

Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O., 2020. nuscenes: A multimodal dataset for autonomous driving, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 11621–11631.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-end object detection with transformers, in: European conference on computer vision, Springer. pp. 213–229.

Chan, H.P., Guo, M., Xu, C.Z., 2022. Grounding commands for autonomous vehicles via layer fusion with region-specific dynamic layer attention, in: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE. pp. 12464–12470.

Chen, X., Ma, L., Chen, J., Jie, Z., Liu, W., Luo, J., 2018. Real-time referring expression comprehension by single-stage grounding network. arXiv preprint arXiv:1812.03426 .

Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J., 2020. Uniter: Universal image-text representation learning, in: European conference on computer vision, Springer. pp. 104–120.

Cheng, B., Schwing, A., Kirillov, A., 2021. Per-pixel classification is not all you need for semantic segmentation. Advances in Neural Information Processing Systems 34, 17864–17875.

Cheng, W., Yin, J., Li, W., Yang, R., Shen, J., 2023. Language-guided 3d object detection in point cloud for autonomous driving. arXiv preprint arXiv:2305.15765 .

Dai, H., Luo, S., Ding, Y., Shao, L., 2020. Commands for autonomous vehicles by progressively stacking visual-linguistic representations, in: Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16, Springer. pp. 27–32.

Deng, C., Wu, Q., Wu, Q., Hu, F., Lyu, F., Tan, M., 2018. Visual grounding via accumulated attention, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7746–7755.

Deruyttere, T., Collell, G., Moens, M.F., 2020a. Giving commands to a self-driving car: A multimodal reasoner for visual grounding. arXiv preprint arXiv:2003.08717 .

Deruyttere, T., Milewski, V., Moens, M.F., 2021. Giving commands to a self-driving car: How to deal with uncertain situations? Engineering applications of artificial intelligence 103, 104257.

Deruyttere, T., Vandenhende, S., Grujicic, D., Liu, Y., Van Gool, L., Blaschko, M., Tuytelaars, T., Moens, M.F., 2020b. Commands 4 autonomous vehicles (c4av) workshop summary, in: Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16, Springer. pp. 3–26.

Deruyttere, T., Vandenhende, S., Grujicic, D., Van Gool, L., Moens, M.F., 2019. Talk2car: Taking control of your self-driving car. arXiv preprint arXiv:1909.10838 .

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 .

Ding, H., Liu, C., Wang, S., Jiang, X., 2021. Vision-language transformer and query generation for referring segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 16321–16330.

Ding, H., Liu, C., Wang, S., Jiang, X., 2022. Vlt: Vision-language transformer and query generation for referring segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence .

Dong, J., Chen, S., Miralinaghi, M., Chen, T., Labi, S., 2022. Development and testing of an image transformer for explainable autonomous driving systems. Journal of Intelligent and Connected Vehicles 5, 235–249.

Dong, J., Chen, S., Miralinaghi, M., Chen, T., Li, P., Labi, S., 2023a. Why did the ai make that decision? towards an explainable artificial intelligence (xai) for autonomous driving systems. Transportation research part C: emerging technologies 156, 104358.

Dong, J., Long, Z., Mao, X., Lin, C., He, Y., Ji, S., 2021. Multi-level alignment network for domain adaptive cross-modal retrieval. Neurocomputing 440, 207–219.

Dong, Z., Zhang, W., Huang, X., Ji, H., Zhan, X., Chen, J., 2023b. Hubo-vlm: Unified vision-language model designed for human robot interaction tasks. arXiv preprint arXiv:2308.12537 .

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 .

Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A., 2010. The pascal visual object classes (voc) challenge. International journal of computer vision 88, 303–338.

Everingham, M., Winn, J., 2012. The pascal visual object classes challenge 2012 (voc2012) development kit. Pattern Anal. Stat. Model. Comput. Learn., Tech. Rep 2007, 5.

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

Grujicic, D., Deruyttere, T., Moens, M.F., Blaschko, M.B., 2022. Predicting physical world destinations for commands given to self-driving cars, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 715–725.

Hao, S., Lee, D.H., Zhao, D., 2019. Sequence to sequence learning with attention mechanism for short-term passenger flow prediction in large-scale metro system. Transportation Research Part C: Emerging Technologies 107, 287–300.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

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

Hudson, D.A., Manning, C.D., 2018. Compositional attention networks for machine reasoning. arXiv preprint arXiv:1803.03067 .

Jain, K., Chhangani, V., Tiwari, A., Krishna, K.M., Gandhi, V., 2023. Ground then navigate: Language-guided navigation in dynamic scenes, in: 2023 IEEE International Conference on Robotics and Automation (ICRA), IEEE. pp. 4113–4120.

Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., Carion, N., 2021. Mdetr-modulated detection for end-to-end multi-modal understanding, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1780–1790.

Kheiri, K., Karimi, H., 2023. Sentimentgpt: Exploiting gpt for advanced sentiment analysis and its departure from current machine learning. arXiv preprint arXiv:2307.10234 .

Li, G., Qiu, Y., Yang, Y., Li, Z., Li, S., Chu, W., Green, P., Li, S.E., 2022a. Lane change strategies for autonomous vehicles: a deep reinforcement learning approach based on transformer. IEEE Transactions on Intelligent Vehicles .

Li, J., Li, D., Xiong, C., Hoi, S., 2022b. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, in: International Conference on Machine Learning, PMLR. pp. 12888–12900.

Li, Z., Chen, Z., Li, Y., Xu, C., 2023a. Context-aware trajectory prediction for autonomous driving in heterogeneous environments. Computer-Aided Civil and Infrastructure Engineering .

Li, Z., Liao, H., Tang, R., Li, G., Li, Y., Xu, C., 2023b. Mitigating the impact of outliers in traffic crash analysis: A robust bayesian regression approach with application to tunnel crash data. Accident Analysis & Prevention 185, 107019.

Li, Z., Xu, C., Bian, Z., 2022c. A force-driven model for passenger evacuation in bus fires. Physica A: Statistical Mechanics and its Applications 589, 126591.

Liao, H., Li, Z., Shen, H., Zeng, W., Li, G., Li, S.E., Xu, C., 2023. Bat: Behavior-aware human-like trajectory prediction for autonomous driving. arXiv preprint arXiv:2312.06371 .

Liao, Y., Liu, S., Li, G., Wang, F., Chen, Y., Qian, C., Li, B., 2020. A real-time cross-modality correlation filtering method for referring expression comprehension, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10880–10889.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, Springer. pp. 740–755.

Loshchilov, I., Hutter, F., 2016. Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 .

Lu, J., Yang, J., Batra, D., Parikh, D., 2016. Hierarchical question-image co-attention for visual question answering. Advances in neural information processing systems 29.

Luo, S., Dai, H., Shao, L., Ding, Y., 2020. C4av: learning cross-modal representations from transformers, in: Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16, Springer. pp. 33–38.

Mittal, V., 2020. Attngrounder: Talking to cars with attention, in: Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16, Springer. pp. 62–73.

OpenAI, 2023. Gpt-4 technical report. arXiv:2303.08774.

Othman, K., 2021. Public acceptance and perception of autonomous vehicles: a comprehensive review. AI and Ethics 1, 355–387.

Ou, J., Zhang, X., 2020. Attention enhanced single stage multimodal reasoner, in: Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16, Springer. pp. 51–61.

Qi, D., Su, L., Song, J., Cui, E., Bharti, T., Sacheti, A., 2020. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. arXiv preprint arXiv:2001.07966 .

Redmon, J., Farhadi, A., 2018. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 .

Reimers, N., Gurevych, I., 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084 .

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems 28.

Rufus, N., Jain, K., Nair, U.K.R., Gandhi, V., Krishna, K.M., 2021. Grounding linguistic commands to navigable regions, in: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE. pp. 8593–8600.

Rufus, N., Nair, U.K.R., Krishna, K.M., Gandhi, V., 2020. Cosine meets softmax: A tough-to-beat baseline for visual grounding, in: Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16, Springer. pp. 39–50.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 .

Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J., 2019. Vl-bert: Pre-training of generic visual-linguistic representations. arXiv preprint arXiv:1908.08530 .

Tan, H., Bansal, M., 2019. Lxmert: Learning cross-modality encoder representations from transformers. arXiv preprint arXiv:1908.07490 .

Tan, M., Le, Q., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks, in: International conference on machine learning, PMLR. pp. 6105–6114.

Tang, D., Qin, B., Liu, T., 2015. Learning semantic representations of users and products for document level sentiment classification, in: Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 1: long papers), pp. 1014–1023.

Vandenhende, S., Deruyttere, T., Grujicic, D., 2020. A baseline for the commands for autonomous vehicles challenge. arXiv preprint arXiv:2004.13822 .

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems 30.

Wang, L., Li, Y., Huang, J., Lazebnik, S., 2018. Learning two-branch neural networks for image-text matching tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence 41, 394–407.

Wang, P., Wu, Q., Cao, J., Shen, C., Gao, L., Hengel, A.v.d., 2019. Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1960–1968.

Wen, K., Xia, J., Huang, Y., Li, L., Xu, J., Shao, J., 2021. Cookie: Contrastive cross-modal knowledge sharing pre-training for vision-language representation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2208–2217.

Yang, K., Lee, D., Whang, T., Lee, S., Lim, H., 2019a. Emotionx-ku: Bert-max based contextual emotion classifier. arXiv preprint arXiv:1906.11565 .

Yang, L., Xu, Y., Yuan, C., Liu, W., Li, B., Hu, W., 2022. Improving visual grounding with visual-linguistic verification and iterative reasoning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9499–9508.

Yang, Z., Chen, T., Wang, L., Luo, J., 2020. Improving one-stage visual grounding by recursive sub-query construction, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16, Springer. pp. 387–404.

Yang, Z., Gong, B., Wang, L., Huang, W., Yu, D., Luo, J., 2019b. A fast and accurate one-stage approach to visual grounding, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4683–4693.

Yu, L., Lin, Z., Shen, X., Yang, J., Lu, X., Bansal, M., Berg, T.L., 2018. Mattnet: Modular attention network for referring expression comprehension, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1307–1315.

Zhuang, B., Wu, Q., Shen, C., Reid, I., Van Den Hengel, A., 2018. Parallel attention: A unified framework for visual object discovery through dialogs and queries, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4252–4261.