Language Explanations for Self-Driving Scenes

1st Vamsi Krishna Sivakavi
dept. Computer Science & Engineering
California State University, San Bernardino
San Bernardino, USA
008458886@coyote.csusb.edu

1st Calvin Glisson
dept. Computer Science & Engineering
California State University, San Bernardino
San Bernardino, USA
008140867@coyote.csusb.edu

1st Alexis Luevanos
dept. Computer Science & Engineering
California State University, San Bernardino
San Bernardino, USA
Luevanos353@gmail.com

4th Qiuxiao Chen
dept. Computer Science & Engineering
California State University, San Bernardino
San Bernardino, USA
chenqiuxiao.ee@gmail.com

***Abstract*—Recent advancements in autonomous driving technology, driven by innovations in computer vision, machine learning, and sensor fusion, have made significant progress. However, interpreting complex urban environments remains a substantial challenge. Dense Cityscapes, with their dynamic interactions between vehicles, pedestrians, cyclists, and varied infrastructure, often confound current models, leading to errors in decision-making. To address these challenges, we explore the integration of natural language processing (NLP) with visual data to enhance the interpretability and safety of autonomous driving systems. This paper presents two novel datasets containing natural language captions: one for the Cityscapes dataset, comprising real-world urban driving scenes, and another for the GTA5 dataset, which features synthetic driving environments capable of simulating rare or hazardous situations. First, using Google's Gemini 1.5 Flash model, we automatically generate scene descriptions focused on traffic-related elements such as vehicle positioning, traffic signals, and pedestrian presence. While machine generated captions provide a baseline, our results show they often misinterpret object positions, traffic flow, and color recognition. In contrast, our human-corrected captions offer more precise and contextually accurate descriptions. The combination of synthetic and real-world data enhances the robustness of autonomous systems, particularly in handling rare and challenging scenarios. Our research advances vision-language integration, with the ultimate goal of improving the interpretability, performance, and safety of autonomous vehicles in urban environments.**

***Index Terms*—text prompt, image caption, dataset, autonomous driving**

I. Introduction

Autonomous driving technology has seen remarkable advances in recent years, fueled by progress in computer vision, machine learning, and sensor fusion. Yet, despite these strides, autonomous vehicles continue to face significant challenges when interpreting complex urban environments. These systems often struggle with dynamic driving scenarios, leading to difficulties in decision-making, particularly in densely populated city scenes where vehicles, pedestrians, cyclists, and varied infrastructure intersect.

Natural language processing (NLP) offers a promising solution to these challenges by pairing visual data with natural language descriptions. This integration can improve the explainability and interpretability of autonomous systems, potentially enhancing safety across diverse driving contexts. By providing context-aware descriptions, this approach can assist both vehicles and human operators in making better-informed decisions.

Urban driving environments, characterized by their complexity and constant flux, present unique challenges. Dense populations of vehicles, pedestrians, and cyclists, along with traffic signals and signage, create an ever-evolving landscape that is difficult for current models to fully comprehend. In this paper, we propose a novel natural language dataset built upon the widely used Cityscapes [1] dataset, renowned for its comprehensive depiction of urban driving scenes captured under various weather conditions and times of day. Our dataset pairs each image with detailed, context-aware textual descriptions, enabling advances in scene comprehension, model interpretability, and autonomous vehicle safety in urban settings.

Additionally, we augment this dataset with a supplementary natural language resource derived from the GTA5 [2] dataset, a synthetic environment offering high fidelity and a broad range of driving scenarios. The GTA5 dataset excels in simulating rare or hazardous situations that are difficult to capture in real-world data. This synthetic dataset is critical in addressing the "long tail" of autonomous driving challenges, including unexpected pedestrian behavior, sudden vehicle maneuvers, and adverse weather conditions. By incorporating both real-world and synthetic data, we aim to improve the robustness and generalizability of autonomous driving systems, especially in handling safety-critical corner cases.

In summary, our contributions are twofold: first, the creation of a comprehensive natural language dataset aligned with Cityscapes, and second, the augmentation of this dataset with synthetic driving scenarios from GTA5. Together, these resources are expected to drive forward research in vision-language integration, with the ultimate goal of enhancing the safety, interpretability, and performance of autonomous vehicles in increasingly complex urban environments.

II. Method

Dataset Introduction.

GTA5 is a synthetic dataset containing 24966 images with per-pixel semantic annotations. Each image is rendered from within the video game Grand Theft Auto 5, from the perspective of a car driving in an American environment. Cityscapes, on the other hand, is an autonomous driving dataset with data collected from 50 cities in Germany. The scenes capture several months of data throughout the spring, summer, and fall seasons and have both pixel-level and instance-level semantic labels. Specifically, there are 5000 Cityscapes images with fine annotations. Recent research has used the GTA5 dataset in conjunction with Cityscapes to perform domain adaptation from the synthetic environment to the real world, obtaining promising results. However, neither dataset provides natural language captions for any of the images.

Dataset Generation. For automatic generation of scene descriptions, we employ Google's Gemini 1.5 Flash model, a multimodal LLM with the ability to interpret image inputs. For each image of each dataset, we generate a caption by providing Gemini with the image and the following prompt: "Describe the image according to traffic data in 2-3 sentences." This encourages the model to focus on describing image information relevant to driving. The resultant captions describe details such as the number of cars, their positions on the road, and the colors of upcoming traffic lights. We note that many of the automatically generated captions have errors, particularly in confusing whether an object is on the left or right side of an image. Therefore, we manually inspect the images and captions to correct any mistakes made by the LLM.

III. Main Results

The machine generated captions reveal significant challenges in accurately identifying the location of vehicles, understanding the direction of traffic, and often providing incomplete or incorrect details. They struggle with identifying the colors of cars and distinguishing between left and right, leading to confusion around object positioning, particularly when distant objects or those at the corners of images are involved. Additionally, the machine generated captions show inconsistency in recognizing the colors of traffic signal lights (e.g., orange, red, or green) and often use inconsistent wording in sentences. In contrast, the human corrected and checked versions of these descriptions are more refined and specific, accurately depicting the scene with clear references to the direction of traffic, the exact placement of vehicles, and the behavior of traffic signals. These revised descriptions also capture additional contextual details, such as the presence of pedestrians and the overall busyness of intersection, elements that the machine often overlooks or misinterprets. For example, in Figure 1 of the GTA dataset image, the machine generated caption states, "A red car is on the right, just past the lane markings, and is not visible in the photo." This is both inaccurate and misleading. The human answer, however,

GTA:

Machine Generated Caption: A white car is driving down a street with a single lane in each direction. A black car is on the left side of the road. A red car is on the right, just past the lane markings, and is not visible in the photo. There is a line of parked cars on the right side of the road and a line of parked cars across the street from them. The white car is driving down the left lane.

Human Corrected Caption: A white car is driving in the right lane of a two-way street, with a single lane in each direction. Several cars are parked along both sides of the road, with three on the left side and two on the right. The road is otherwise clear of moving traffic, and there is a fire hydrant visible on the left side.

Cityscapes:

Machine Generated Caption: The image shows a street with a bus driving down the center lane. There are cars parked on the right side of the street. The bus is driving towards the left side of the image, so the road is likely one way. The white lines indicate the lane divisions. The car in the foreground is also driving down the center lane.

Human Corrected Caption: The image shows a street with two buses driving down the right lane toward the stoplight. Seven cars are parked on the right side of the street, and one car is parked on the left side. The buses are driving towards the right side of the image, indicating that the road is likely a multi-lane road with clearly marked lanes for traffic in both directions. The white lines indicate the lane divisions, and the car in the foreground is also driving down the center lane.

Fig. 1. Two sample images and their machine generated captions alongside the human corrected captions are presented. The Text Prompt for these images is: "Describe the image according to traffic data in 2-3 sentences."

accurately states, "Several cars are parked along both sides of the road, with three on the left side and two on the right," offering a more comprehensive and precise description. This highlights the machine's difficulty in accurately interpreting complex traffic scenes, while the human response captures all necessary details. Moreover, small changes in structure and punctuation can significantly enhance the clarity and impact of a description, making human corrected captions far more effective in conveying accurate and detailed information than the machine generated captions.

REFERENCES

[1] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[2] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer, 2016, pp. 102–118.