DeepRide: Dashcam Video Description Dataset for Autonomous Vehicle Location-Aware Trip Description

GHAZALA RAFIQ, MUHAMMAD RAFIQ, BYUNG-WON ON, MANKYU SUNG, AND GYU SANG CHOI, (Member, IEEE)

1Department of Information and Communication Engineering, Yeungnam University, Gyeongsan-si 38541, South Korea
2Department of Game and Mobile Engineering, Keimyung University, Daegu 42601, South Korea
3Department of Software Convergence Engineering, Kunsan National University, Gunsan-si 54150, South Korea

Corresponding authors: Gyu Sang Choi (castchoi@ynu.ac.kr) and Muhammad Rafiq (rafiq@kmu.ac.kr)

This work was supported in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant NRF-2021R1A6A1A03039493, Grant NRF-2018R1D1A1B07048414, and Grant 2022R1A2C1011404; and in part by the 2022 Yeungnam University Research Grant.

ABSTRACT Video description is one of the most challenging tasks in the combined domain of computer vision and natural language processing. Captions for various open and constrained domain videos have been generated in the recent past but descriptions for driving dashcam videos have never been explored to the best of our knowledge. With the aim to explore dashcam video description generation for autonomous driving, this study presents DeepRide: a large-scale dashcam driving video description dataset for location-aware dense video description generation. The human-described dataset comprises visual scenes and actions with diverse weather, people, objects, and geographical paradigms. It bridges the autonomous driving domain with video description by textual description generation of the visual information as seen by a dashcam. We describe 16,000 videos (40 seconds each) in English employing 2,700 man-hours by two highly qualified teams with domain knowledge. The descriptions consist of eight to ten sentences covering each dashcam video’s global features and event features in 60 to 90 words. The dataset consists of more than 130K sentences, totaling approximately one million words. We evaluate the dataset by employing location aware vision-language recurrent transformer framework to elaborate on the efficacy and significance of the visio-linguistics research for autonomous vehicles. We provided baseline results to evaluate the dataset by employing three existing state-of-the-art recurrent models. The memory augmented transformer performed superior due to its highly summarized memory state for visual information and the sentence history while generating the trip description. Our proposed dataset opens a new dimension of diverse and exciting applications, such as self-driving vehicle reporting, driver and vehicle safety, inter-vehicle road intelligence sharing, and travel occurrence reports.

INDEX TERMS Dashcam video description, video description dataset, video captioning, autonomous trip description.

I. INTRODUCTION

Automatic description generation is an established and challenging task for short as well as relatively long videos. It has achieved intense attention in the realm of computer vision, and natural language processing [1], exploring a variety of constrained and open domains. Developing robust video description systems [2], [3], [4], [5], [6], [7] demands not only the ability to understand the sequential visual data but also to generate a syntactically concise and semantically accurate interpretation of that understanding into natural language. The accomplishment of accurate and diverse description
generation is directly associated with the amount and quality of the training and validation data provided to the model for understanding.

Comprehending the localized events of a video appropriately and then transforming the attained visual understanding accurately into a textual format is called dense video captioning, or simply, video description [8]. Capturing the scenes, objects, and activities in a video, as well as the spatial–temporal relationships and the temporal order, is crucial for precise and grammatically correct multi-line text narration. The generated fine-grained caption is a requirement of such a mechanism that proves to be expressive and subtle. Its purpose is to bag the temporal dynamics of the visuals in specific order as presented in the video, and then join them with syntactically and semantically correct representations using natural language.

Considering the training-test data for video description systems, various datasets have been proposed in the recent past for better visual comprehension and diverse description generation. These datasets belong to a variety of domains, discovering aspects associated with our daily lives like human actions [9], [10], cooking [11], [12], [13], [14], [15], [16], movies [17], [18], social media [19], [20], TV shows [21], E-commerce [22] and generalized categories [8], [23], [24], [25].

A. MOTIVATION

The emerging autonomous vehicle technology has achieved increasing attention in the recent past. There is a great deal of research being conducted in various research sectors of autonomous driving, particular to computer vision tasks; object detection, semantic segmentation, semantic instance segmentation and depth estimation are some of them. However, video description for driving videos has never been explored to the best of our knowledge. Blending the challenging video description domain with the promising autonomous driving research can definitely push the frontiers of the research in an ambitious direction. With the motivation to understand the challenges of video description system in the context of autonomous driving, we collect a novel, large-scale, location-aware dashcam videos description dataset, DeepRide. The proposed dataset features 16k dashcam videos corresponding to more than 130k sentences in 16k paragraph descriptions. Each description on average has ten sentences describing the day/night time, weather information, scene attributes along with static features and dynamic events. Static features include parked cars, trees, signboards, and high-rise buildings on the road side whereas, by dynamic events we mean the switching of traffic signals at the intersection, turning of vehicle, passing under/over a bridge, wiping windshield and an accident happening. Exploration of the challenging video description research applied in the potentially exciting domain of autonomous driving, no doubt, represents an expanded challenge in this research area.

B. DEEPRIIDE APPLICATIONS

The importance of video description is evident from its practical and real-time applications, i.e., efficient searching and indexing of videos on the internet, human-robot relationships in industrial zones, facilitation of autonomous vehicle driving. Video descriptions can outline procedures in instructional/tutorial videos for industry, education, and the household (e.g., recipes). The visually impaired can gain useful information from a video that incorporates audio descriptions. Long surveillance videos can be transformed into short texts for quick previews. Sign language videos can be converted to natural language descriptions. Automatic, accurate, and precise video/movie subtitling is another important and practical application of the video description task.

Particular to DeepRide dataset, the basic application or the purpose of dataset creation is to automatically generate summaries (trip descriptions) for autonomous vehicles using dashcam videos. The desired generated summary contains the vehicle’s location information from the GPS data stored by default in the dashcam meta data, day/night, weather, scene, road side information (trees, buildings, parking), and dynamic events taking place on and around the road (vehicle position and speed on the road, traffic signals, turnings, entering/exiting underpass/overheads, traffic flow, pedestrians movements/waiting, accident occurring). Other significant and noteworthy applications can be

- self-driving vehicle reporting
- driver and vehicle safety
- inter-vehicle road intelligence sharing
- travel occurrence reports.

C. LOCATION-AWARE DESCRIPTION

In the desire to get human-like precise and accurate descriptions for a supplied video, various strategies have been investigated for quality enhancement and optimization. Since a video naturally comprises of multiple modalities, i.e., visual, audio, sound, and sometimes subtitles, so catering the available modes within a video can result in accelerated training and boosted performance. Dealing with trip descriptions from dashcam videos, one of the important aspects is the location, i.e., embedded GPS/IMU information recorded automatically with the visual data. A trip summary is considered incomplete without the location information so utilization of GPS data can help in vehicle’s location detection. We propose a location-aware recurrent transformer based dashcam video description framework for the generation of rich and informative trip description. Deeming the location-aware feature, Our proposed dataset opens a new dimension of diverse and exciting applications stated above.

Our contributions towards this research work are as follows:

1) We explore a new direction for the task of video description by blending it with the fast-growing and emerging domain of autonomous vehicle driving.
2) We collect a novel, large scale, location-aware video description dataset using dashcam videos for autonomous vehicle trip description.

3) We employ a state-of-the-art web-based platform for the systematic collection, revision, proofreading, and finalization of the data.

4) We perform in-depth analysis of the collected data and further investigate the shared recurrent transformer with the location framework for the generation of natural language descriptions and validate the system’s efficiency and effectiveness.

D. PROBLEM STATEMENT

Assume that we have a dashcam video \( V \), containing multiple temporal event sections \( \{E_1, E_2, E_3, \ldots E_T\} \) and GPS/IMU information. An automatic sentence is generated for every event in the video to describe the content thereof in natural language. Our goal is to come up with a location-aware coherent multi-sentence \( \{S_0, S_1, S_2, S_3, \ldots S_T\} \) description, where \( T \) represents number of events and \( S_1 \) through \( S_T \) are the corresponding generated sentences, whereas \( S_0 \) is the first sentence of generated description providing location information. Figure 1 demonstrates some sample dashcam video frames from training set of DeepRide dataset with ground truth description.

Rest of the paper is organized as follows: Section II provides a brief overview of the related literature on the topic, Section III explores the collection, statistics and analysis of DeepRide dataset followed by Section IV presenting the proposed multi-modal location-aware recurrent transformer based video description framework. Section V elaborates the experimentation and implementation details. The Qualitative and quantitative results are presented in Section VI and finally, the paper is concluded in Section VII with few future directions.

II. RELATED WORK

Datasets creation related to computer vision tasks have played significant part in developing algorithms for robust performance. Creating broadly challenging and ambitious datasets
TABLE 1. Benchmark multi-caption video description datasets.

| Data set      | Domain | Videos | Duration (Hours) | Descriptions | MC  |
|---------------|--------|--------|------------------|--------------|-----|
| MSVD          | General| 1,970  | 5.3              | 70,028       | N   |
| VATEX         | General| 41,269 | -                | 412,690      | N   |
| YouCookII     | Cooking| 2k     | 17.6             | 15.4k        | Y   |
| ANet Captions | Multi-category| 20k | 849              | 100k         | Y   |
| LoL-V2T       | Video-Game| 9.7k | 76               | 63k          | Y   |

DeepRide (ours) | Driving| 16k | 177h | 130K | Y   |

MC: Multi-Sentence Captions (paragraph like description) associated with each video. Y: Yes, N: No

TABLE 2. DeepRide dataset split (train, validation, and test) statistics.

| Split       | Videos | Sentences | Avg Sent/Desc | Words | Avg Words/Desc |
|-------------|--------|-----------|---------------|-------|---------------|
| Train       | 11k    | 92.4k     | 9.2           | 670k  | 68            |
| Validation  | 3.5k   | 29.4k     | 9.3           | 213k  | 69            |
| Test        | 1.5k   | 12.6k     | 9.1           | 91k   | 67            |
| Total       | 16k    | 134.4k    | -             | 974k  | -             |

Avg: Average, Sent: Sentence, Desc: Description

can take vision-to-language research in a distinct direction and provide organized means for training and evaluations. The publicly available datasets with deep and diverse descriptions, novel tasks and challenges, and meticulous benchmarks have contributed intensely for recent rapid developments in the Visio-linguistic field. The intersection of computer vision for autonomous driving with natural language processing by [26], [27], [28], and [29] is pushing forward the frontiers of research domain in a new direction altogether.

A. VIDEO DESCRIPTION DATASETS

Various datasets have been launched from time to time to exhibit an enhanced accomplishment for the task of video description, exploring a wide range of constrained and open domains like cooking by [11], [12], [13], [14], [15], and [16], human activities by [8], [9], [23], [24], and [25], social media by [20], and [19], movies by [17], and [18], TV shows by [21], and e-commerce by [22] presented in detail by [30]. Table 1 lists a brief overview of the key attributes and major statistics of existing multi-caption (dense/paragraph like) video description datasets. The existing renowned datasets gradually heightened their visual complexity and language diversity to expand dynamic and hefty algorithms.

B. VIDEO DESCRIPTION APPROACHES

Video description generation approaches can be broadly classified into four groups based on their technological advancement in time.

1) **Encoder-Decoder (ED) based Approaches**: The ED framework is the most popular paradigm for video description generation [31], [32], [33], [34], [35] in recent years, it pioneered the video description task by addressing the limitations associated with conventional and statistical approaches. Conventional ED pipelines typically consist of a CNN being used as a visual model to extract visual features from video frames, and an RNN being used as a language model to generate captions word by word. Other compositions of CNN, RNN and their variants LSTMs and GRUs are also explored in this field following the ED architecture.

2) **Attention Mechanism based Approaches**: The standard encoder-decoder architecture further fused with attention mechanism to focus on specific distinctness showed high quality performances. The captioning system developed by [36], [37], [38], [39], [40], [41], and [42] demonstrated the employment of visual, local, global, adaptive, spatial, temporal, and channel attention for coherent and diverse caption generation.

3) **Transformer based Approaches**: Recently with the advent of efficient and modern transductive transformer architecture, free from recurrence, and solely based on self-attention, video description systems enhanced the performance allowing parallelization along with training on massive amount of data. With the emergence of several versions of transformers and models employing transformers [2], [3], [4], [5], [43], [44], [45], [46], [47] long term dependency handling is not an issue anymore for researchers engaged in video processing for summarization and description, or for autonomous-vehicle, surveillance, and instructional purposes.

4) **Deep Reinforcement Learning based Approaches**: Reinforcement learning employed within the encoder-decoder structure [48], [49], [50], [51], [52] can progressively deliver state-of-the-art captions by following exploration and exploitation strategies. Recently, the notion of deep reinforcement learning in the video description domain with the capacity of repeated polishing [53] simulates human cognitive behaviors.
The proposed model-irrelevant algorithm introduced a polishing mechanism into the video description via reinforcement learning and gradually improved the generated captions by revising the ambiguous word and grammar errors.

C. DRIVING DATASETS

Domain-specific, large-scale, and diverse datasets can fuel further advances in supervised learning. In the fast-growing field of autonomous driving, the datasets BDD-100K [26], NuScenes [27], KAIST multi-spectral driving dataset [28], KITTI [29], ROAD [54], and A2D2 [55] have proven to be of great value for computer vision tasks like object classification, object detection, and scene segmentation.

BDD-100K dataset [26] consists of 100K video clips embracing realistic driving scenarios with increasing complexities for heterogeneous multitask learning. The crowd sourced dataset, solely collected from drivers, explored for ten computer vision tasks involving image and tracking tasks.

Through this research work, associating the renowned area of autonomous driving with the video description domain can exclusively proceed the research with a distinct focus.

III. DATASET: DEEPRIDE

This section presents the videos collection, descriptions collection, description collection framework, Data batches, and statistics of the DeepRide dataset.

A. VIDEOS COLLECTION

DeepRide dataset is created with the objective to discover the challenging video description task in conjunction with the emerging autonomous driving domain. The 16k dashcam videos encompassing diverse driving scenarios are taken from BDD100K [26], the large-scale driving video dataset exposing the challenges of street-scene understanding. The dashcam videos, each of 40 seconds duration, are obtained in a crowd-sourcing manner from more than 50K rides, primarily uploaded by thousands of drivers covering New York, Berkeley, San Francisco, Bay Area, and other regions in the populous areas of the USA and around the world. Vehicle’s dashcam is used to record these videos with 30-fps frame rate along with GPS/IMU information preserving the driving trajectories. The GPS/IMU information is employed to generate location-aware descriptions. These videos are recorded at different times of the day with diverse weather conditions and varied scene locations. The dashcam video’s three global features/characteristics given below are also considered for the natural language descriptions generation.

1) Time like dawn/dusk, day, and night time (sample frames shown in Figure 9)
2) Weather conditions including rainy, snowy, foggy, overcast, cloudy, and clear (sample frames shown in Figure 10)
3) Scene type such as residential area, city-street, and highways (sample frames shown in Figure 11)
Appreciating the original train-test split of the BDD100k dataset, among 16k dashcam videos of our DeepRide, 11k videos are in the training set, taken from the training split of BDD100k and 5k videos from the validation set of BDD100k constitutes the validation and test set of our collected dataset as shown in the Table 2.

B. DESCRIPTION COLLECTION FRAMEWORK

We designed and developed a proprietary web-based portal for the descriptions collection of 16k dashcam videos. Data entry operators are examined for their driving knowledge and road experience by conducting basic test and interviews. The 75% of operator qualification added value to the data quality. The qualified operators are assigned to describe each dashcam video in eight to ten concise yet descriptive sentences (not limited to ten sentences, if there is more to describe, they can) covering all the static scenes and dynamic events taking place on and around the road. Static scenes include parked cars, trees, signboards, and high-rise buildings on the roadside, whereas, by dynamic events, we mean the switching of traffic signals at the intersection, turning of the vehicle, passing under/over a bridge, an accident happening, and wiping windshield. An overview of the DeepRide description collection procedure is demonstrated in Figure 2.

In order to ensure smooth operation at the operator’s end, the videos of dashcams have been adjusted for their high definition (HD) quality. Multiple role-specific screens are available in the portal to deal with basic entry, revision, proof reading, and finalization parallel to administration dashboards and description entry statistics screens. These screens include basic description entry screen shown in Figure 3.
revision screen shown in Figure 4 and dashboard screens for administrators shown in Figure 6.

We constitute two teams of highly qualified English-speaking operators with domain knowledge, first for video primary description and second for spelling, grammar, quality check, and proofreading of the primary descriptions. We instruct description operators for concise yet descriptive eight to ten sentences descriptions for each dashcam video in English.

The web portal groups 100 × dashcam videos selected from the dashcam videos pool to constitute a batch. The batch is assigned (batch status: Assigned) to a specific operator by the administrator. On completion of batch description, the operator submits (batch status: Submitted) the batch back to the administrator for revision (batch status: Revision). The batch is checked for spelling, grammar, and description quality. The batch is assigned back to the same operator for corrections if it does not satisfy the description standards. Operators with more than 10% rejections disqualify for further description tasks. Upon acceptance, the batch is further assigned to the proofread operator (batch status: Proofread) for proofreading purposes. The proofreading operator checks for spelling, grammar, and description quality. After proofreading, the batch is pushed to administrator section for batch finalization (batch status: Finalized). The batch statuses are described in Table 3.

C. DATASET STATISTICS
DeepRide is a dense dashcam videos description dataset spanning 177 hours with 16k paragraphs having vocabulary density of 0.004 and readability index of 5.209. Our dataset have a paragraph of eight to ten diverse temporally described sentences for every dashcam video. It makes up to 976,941 total words with 3,722 unique words resulting in overall 130K sentences. The average description length of nine sentences with 68 words on average makes it superior as compared to other datasets shown in Table 1. We layout detailed statistics in Table 2. The Figure 5 represents English word cloud for the DeepRide dataset (top 200 most frequent words).

IV. METHOD
We developed a location-aware video description evaluation framework that generates human analogous descriptions for dashcam videos. We employ various transformer based dense video captioning models to evaluate our proposed dataset. An overview of the proposed framework is shown in
The augmented memory block leverages the video segments and their previous caption history to assist with next sentence generation. We generate our dataset corpus compliant with the ActivityNet-Captions dataset in JavaScript Object Notation (JSON) file format. We evaluate our dataset with metrics of BLEU (1 to 4), CIDEr, ROUGE L, METEOR, and Repetition (1 to 4). We investigate the results of the following models while evaluating our proposed dataset DeepRide.
A. MASKED TRANSFORMER
Considering Neural Machine Translation (NMT), implementing a self-attention mechanism with the objectives of parallelization, reduction of computational complexity, long-range dependency handling [44] introduced basic transformer architecture, which was employed for the video to text paragraph description generation by [46]. They proposed a masking network comprising a video encoder, proposal decoder and a captioning decoder aiming to decode the proposal-specific representations into differentiable masks, resulting in consistent training of proposal generation and captioning decoder. Learning the representation with the capability of long-range dependencies is addressed by employing self-attention, facilitating more effective learning.

B. TRANSFORMER-XL
Introducing the notion of recurrence into pure self-attention based networks, transformer-XL [47] is capable of a paragraph-like description generation by learning beyond a fixed and without interrupting the temporal coherence. They introduced simple yet effective position encoding to generalize attention weights beyond training and the reusability of hidden states to build up the recurrent connection between the segments.

C. MART
The Mart proposed by [5] for the video to text paragraph description generation task is based on vanilla transformer model [44]. Unlike the vanilla model with separate encoder-decoder networks, mart introduced a shared encoder-decoder environment with an auxiliary memory module to enable recurrence in transformers. The augmented external memory block similar to LSTM [56] and GRU [57] facilitates the processing of caption history information corresponding to video segments. The shared environment of encoder-decoder and implementation of memory module by MART allow it to utilize previous contextual information so that it is able to produce a better paragraph that is more coherent and less repetitive.

D. LOCATION AWARE DESCRIPTION GENERATION
The proposed dataset DeepRide utilizes the GPS/IMU recording of preserved trajectory information while processing the corresponding dashcam video to generate a location-aware road trip description. The Latitude and Longitude associated with the dashcam videos are cached from Google Geocoding API with their corresponding position/location containing road & city name and stored in a geographic repository. This database is utilized to get the location associated with the latitude and longitude of the dashcam video while generating trip summary.

Further, the location containing sentence is concatenated with the generated paragraph summary to form location-aware trip description as demonstrated in Figure 8.

V. EXPERIMENTATION
A. FEATURE EXTRACTION
In order to keep the scenes standardized/uniform and get features, we sample 15 frames per second and extract I3D features [58] from these sampled frames. The sampling mechanism is based on time not on frame rate. If a dashcam video is either 30 or 60 fps encoded, the trip description system will sample and process 15 frames per second. If frame rate is less than the required frames/second, i.e. 15 fps, the system will achieve the desired frames per second by adding zero padding.

We feed 64 frames with a spatial size of $224 \times 224$. For better feature representations, we use the I3D model, pre-trained on the Kinetics training dataset [59] and calculated video RGB and optical flow features prior to the training. We extract the temporal features using PWCNet by [60]. The I3D, the spatial/RGB 1024D feature vectors, and temporal/optical flow 1024D feature vectors are concatenated to form the input to the transformer layers. It formed a single 2048D representation for every stack of 64 frames. The dashcam videos are 40 seconds long, hence make ten segments of $244 \times 224 \times 64$, which is sufficient to generate ten sentences.

We employ Glove-6B, 300 dimension word embeddings and generated vocabulary index for language model.
TABLE 5. Description results on test set of DeepRide dataset. ↑ follows the Higher the better, whereas ↓ follows lower the better policy.

| Model          | BLEU@4↑  | CIDEr↑   | METEOR↑ | ROUGE-L↑ | Re@4↓   |
|----------------|---------|---------|---------|---------|---------|
| Masked-Transformer | 9.10    | 5.35    | 14.57   | 29.61   | 77.47   |
| Transformer-XL  | 26.34   | 40.35   | 23.25   | 39.96   | 25.54   |
| MART            | 30.59   | 51.97   | 25.88   | 42.21   | 12.48   |

FIGURE 10. Sample frames from the train set representing the weather attribute: clear, overcast, foggy, cloudy, rainy, and snowy.

B. IMPLEMENTATION DETAILS

We adapt implementation details of the MART [5] for coherent video description generation. MART uses two transformers with 12 multi-head attentions, a hidden size of 768, and a positional encoding based as described in [44]. A memory module with a recurrent memory state of one is included in the model.

A major challenge in training a machine learning model is determining how many epochs should be run. Insufficient epochs are likely to delay the model’s convergence, while excessive epochs may represent overfitting. In order to reduce overfitting without compromising model accuracy, early stopping is a technique used to optimize the model. Early stopping as a form of regularization is primarily concerned with stopping training before an over-fitted model occurs. It is used when we don’t want to train our model for too long, saving computational power.

We employed an early stopping with a patience of 10 epochs. With the greedy decoding approach, we measured CIDEr-D as the primary evaluation parameter and early stopping condition. We show simulation parameters in Table 4. We train the model for 50 epochs, using Adam optimizer with five epoch warm up, an initial learning rate of 0.0001 with $\beta_1$ of 0.9, $\beta_2 = 0.999$, and weight decay of 0.01.

C. EVALUATION METRICS

We evaluate the model using popular automatic evaluation metrics with dense video captioning: Bilingual Evaluation Understudy (BLEU) [61], Consensus-based Image Description Evaluation (CIDEr) [62], Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [63], Metric for Evaluation of Translation with Explicit ORdering (METEOR) [64], and Repetition. We employ standard evaluation sources from MSCOCO server.

VI. RESULTS & DISCUSSION

We compare three transformer-based models and record the results. In Table 5, we reported all three results however observed that only MART based result demonstrated superior performance on DeepRide dataset for all of the evaluation metrics. The other two Transformers: Transformer-XL and Masked Transformer models, demonstrated an average performance. The high performance of MART transformer is fundamentally caused by the memory module, and it has shown a good performance due to the nature of the dataset descriptions.
**FIGURE 12.** Description Analysis - English ground truth description of the sample dashcam videos from train set of DeepRide dataset. Brown-colored sentence is the trip location information. Blue-colored Text indicate Static scenes whereas green-colored text indicate dynamic events taking place on and around the road.

| Ground Truth Description - English |
|-----------------------------------|
| It is 111th Street, NY 11101, New York. The vehicle is moving on a city road. It is daytime. The vehicle turns right. There are buildings on both sides of the road. Cars are parked on both sides of the road. There are trees on both sides of the road. The vehicle is on the central lane of the road. There are signboards on the road. The vehicle collides with a motorcyclist and the windshield breaks. The vehicle stops at the intersection. |

| Ground Truth Description |
|--------------------------|
| It is 3rd Avenue, New York, NY 11220. The vehicle is moving on a city road. The bridge is visible on the left side of the road. Pedestrians are on sidewalks. Shops are on the right side of the road. The vehicle turns to the right side of the road. Lots of cars are parked on both sides of the road. Plenty of trees are visible on both sides. The vehicle moves at a high |

| Ground Truth Description |
|--------------------------|
| It is Namir 136, Tel Aviv-Yafo, Israel. The vehicle speed is high. A petrol pump is nearby. The truck stops on the petrol pump. Power poles are visible on the right side of the road. Vehicles are moving at the front. The vehicle stops behind the bus. Vehicles are moving on the right side of the road. The bridge is visible at the front. |

| Ground Truth Description |
|--------------------------|
| It is Namir 61, 6230231 Tel Aviv-Yafo, Israel. The vehicle is moving on a city road. Plenty of trees are visible on the right side. Lots of vehicles are parked on the left side. A bus moves at the front of the vehicle. The vehicle is in the right most lane of the road. The vehicle moves at a stable speed. High-rise buildings are visible at the front. The vehicle moves at the daytime. The vehicle stops on a red signal. Pedestrians are crossing the road using the zebra crossing. Traffic from the left side is crossing the intersection. |

FIGURE 12. Description Analysis - English ground truth description of the sample dashcam videos from train set of DeepRide dataset. Brown-colored sentence is the trip location information. Blue-colored Text indicate Static scenes whereas green-colored text indicate dynamic events taking place on and around the road.
FIGURE 13. Qualitative Analysis - English ground truth and generated descriptions of the sample dashcam videos from test set of DeepRide dataset. Blue-colored Text indicate Static scenes whereas green-colored text indicate dynamic events. Included only Location-aware MART produced qualitative results.
The driving video datasets are challenging and have a bunch of similar features within every dashcam video. Therefore, there is a significant amount of text description that could repeat due to feature similarities as shown in Figure 9, 10, and 11. The (MART) took advantage of memory and generated far better sequences of sentences while describing the video features. These attributes show various scenes from the dataset, where every scene has typical graphics, i.e., road, marking lines, cars, zebra crossing, signals, building, trees, parked vehicles, pedestrians etc. Therefore, it is quite possible that once a data entry operator describes that vehicle is moving on the road, the model can predict this sentence from every frame because the features are present throughout the video, this sometimes causes the model to predict the sentence at some other time slot since it is a global scenario.

Similarly, the local event features to the timeline are predicted at the time of occurrence, i.e., vehicle stops, turns right, slows down, crosses underpass, pedestrians crossing etc. Although the scene predictions are global and can be listed at any specific time, we have obtained encouraging results, setting a baseline for further improvements. We show description analysis in Figure 12 and qualitative results in Figure 13.

VII. CONCLUSION & FUTURE WORK
In this research work, we present DeepRide, a new diverse location-aware dashcam video description dataset intended to explore emerging autonomous vehicles driving in the perspective of the fast-growing video description domain. Featuring 16k dashcam videos linked with around 130k sentences of description in English. This dataset may help automate the creation of driving commentary. Moreover, the embedded GPS/IMU information recording capability of dashcam video empowers the description system to associate the concerned locations and positions with natural language descriptions. We provided guidelines integrating location information extraction with the recurrent transformers.

Further, our proposed dataset opens a new dimension of diverse and exciting applications: self-driving vehicle reporting, driver and vehicle safety, inter-vehicle road intelligence sharing, and travel occurrence reports. Our future efforts will include creating descriptions for all dashcam videos publicly available by BDD100k, focusing on videos recorded by rear camera, extending the language domain from single to multilingual, along with object detection and relational features research.

We anticipate that the DeepRide dataset’s release will help advance the Visio-linguistic research.

REFERENCES
[1] S. Bhatt, F. Patwa, and R. Sandhu, “Natural language processing (almost) from scratch,” in Proc. IEEE 3rd Int. Conf. Collaboration Internet Comput. (CIC), Jan. 2017, pp. 328–338.
[2] L. Li, Y.-C. Chen, Y. Cheng, Z. Gan, L. Yu, and J. Liu, “HERO: Hierarchical encoder for video+language omni-representation pre-training,” in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), 2020, pp. 2046–2065.
[3] S. Gning, M. Zolfaghari, H. Pirsiavash, and T. Brox, “COOT: Cooperative hierarchical transformer for video-text representation learning,” in Proc. NeurIPS, 2020, pp. 1–27.
[4] T. Jin, S. Huang, M. Chen, Y. Li, and Z. Zhang, “SBAT: Video captioning with sparse boundary-aware transformer,” in Proc. 29th Int. Joint Conf. Artif. Intell., Jul. 2020, pp. 630–636.
[5] J. Lei, L. Wang, Y. Shen, D. Yu, T. Berg, and M. Bansal, “MART: Memory-augmented recurrent transformer for coherent video paragraph captioning,” in Proc. 58th Annu. Meet. Assoc. Comput. Linguistics, 2020, pp. 2603–2614.
[6] Z. Yu and N. Han, “Accelerated masked transformer for dense video captioning,” Neurocomputing, vol. 445, pp. 72–80, Jul. 2021, doi: 10.1016/j.neucom.2021.03.026.
[7] M. Hosseinzadeh and Y. Wang, “Video captioning of future frames,” in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), Jan. 2021, pp. 980–989.
[8] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles, “Dense-captioning events in videos,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 706–715.
[9] D. L. Chen and W. B. Dolan, “Collecting highly parallel data for paraphrase evaluation,” in Proc. 49th Annu. Meet. Assoc. Comput. Linguistics: Hum. Lang. Technol. (ACL-HLT), vol. 1, 2011, pp. 190–200.
[10] G. A. Sigurdsson, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, “Holylow in homes: Crowdsourcing data collection for activity understanding,” 2016, arXiv:1604.01753.
[11] M. Rohrbach and M. Plank, “A database for fine grained activity detection of cooking activities,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2012, pp. 1194–1201.
[12] P. Das, C. Xu, R. F. Doell, and J. J. Corso, “A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2013, pp. 2634–2641.
[13] M. Regneri, M. Rohrbach, D. Wetzel, S. Thater, B. Schiele, and M. Pinkal, “Grounding action descriptions in videos,” Trans. Assoc. Comput. Linguistics, vol. 1, pp. 25–36, 2013.
[14] A. Rohrbach, M. Rohrbach, W. Qiu, A. Friedrich, M. Pinkal, and B. Schiele, “Coherent multi-sentence video description with variable level of detail,” in Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 8753, 2014, pp. 184–195. [Online]. Available: https://arxiv.org/abs/1403.6173.
[15] L. Zhou, C. Xu, and J. J. Corso, “Towards automatic learning of procedures from web instructional videos,” in Proc. 32nd AAAI Conf. Artif. Intell., 2018.
[16] G. Huang, B. Pang, Z. Zhu, C. Rivera, and R. Soricut, “Multimodal pre-training for dense video captioning,” 2020, arXiv:2011.11760.
[17] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele, “A dataset for movie description,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 3202–3212.
[18] A. Torabi, C. Pal, H. Larochelle, and A. Courville, “Using descriptive video services to create a large data source for video annotation research,” 2015, arXiv:1503.01070.
[19] L. Zhou, Y. Kalantidis, X. Chen, J. J. Corso, and M. Rohrbach, “Grounded video description,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 6571–6580.
[20] S. Gella, M. Lewis, and M. Rohrbach, “A dataset for telling the stories of social media videos,” in Proc. Conf. Empirical Methods Natural Lang. Proc., 2018, pp. 968–974.
[21] J. Lei, L. Yu, T. L. Berg, and M. Bansal, “TVR: A large-scale dataset for video-subtitle moment retrieval,” Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (Lecture Notes in Computer Science), vol. 12366. 2020, pp. 447–463. [Online]. Available: https://arxiv.org/abs/2001.09099v2.
[22] S. Zhang, Z. Tan, J. Yu, Z. Zhao, K. Kuang, J. Liu, J. Zhou, H. Yang, and F. Wu, “Poet: Product-oriented video captioner for E-commerce,” 2020, arXiv:2008.09850.
[23] J. Xu, T. Mei, T. Yao, and Y. Rui, “MSR-VTT: A large video description dataset for bridging video and language,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 5288–5296.
[24] K. H. Zeng, T. H. Chen, J. C. Niebles, and M. Sun, “Title generation for user generated videos,” in Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (Lecture Notes in Computer Science), vol. 9906. 2016, pp. 609–625. [Online]. Available: https://arxiv.org/abs/1608.07068
MUHAMMAD RAFIQ (Member, IEEE) received the M.S. degree in electronics engineering from International Islamic University, Pakistan, in 2008, and the Ph.D. degree in information and communication engineering from Yeungnam University, South Korea, in 2022. He has extensive industry experience with a background in databases, business applications, and industrial technology solutions. His research interests include modern 3-D game development, computer vision, video description incorporating artificial intelligence, and deep learning.

Byung-Won On received the M.S. degree from the Department of Computer Science and Engineering, Korea University, Seoul, South Korea, in 2000, and the Ph.D. degree from the Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA, USA, in 2007. He is currently a Professor with the Department of Software Science and Engineering, Kunsan National University, Gunsan-si, South Korea, where he also leads the Data Intelligence Laboratory. His current research interests include around data mining, in particular probability theory and applications, machine learning, and artificial intelligence, mainly working on abstractive summarization, creative computing, and multiagent reinforcement learning. He is currently serving as a Committee Member for ISO/IEC JTC 1/SC 32, Korean Association of Data Science, and SIG on Human Language Technology with the Korean Institute of Information Scientists and Engineers (KIISE), and the Informatization Committee and Jeonbuk Large Leap Policy Consultation Body in Jeollabuk-do Provincial Government. He is an Editor of Journal of Korean Institute of Information Scientists and Engineers (KIISE), Electronics and Telecommunications Research Institute Journal (ETRI), and Quality and Quantity.

MANKYU SUNG received the B.S. degree in computer science from Chungnam National University, Daejeon, in 1993, and the M.S. and Ph.D. degrees in computer science from the University of Wisconsin–Madison, WI, USA, in 2005. From January 1995 to July 2012, he worked for Digital Contents Division, ETRI, Daejeon, South Korea. He has been an Assistant Professor, Department of Game and Mobile, Keimyung University, Daegu, South Korea, since March 2012. His current research interests include computer graphics, deep-learning applications, computer animation, computer games, and human–computer interaction. He is a member of the ACM.

Gyu Sang Choi (Member, IEEE) received the Ph.D. degree in computer science and engineering from Pennsylvania State University. He was a Research Staff Member at the Samsung Advanced Institute of Technology (SAIT), Samsung Electronics, from 2006 to 2009. Since 2009, he has been with Yeungnam University, where he is currently an Assistant Professor. He is currently working on embedded systems and storage systems, while his prior research has been mainly focused on improving the performance of clusters. His research interests include embedded systems, storage systems, parallel and distributed computing, supercomputing, cluster-based web servers, and data centers. He is a member of the ACM.

* * *