Review of DonkeyCar Platform: From Deep Learning to Autonomous Driving Implementation

Ning Wang[1] and Jian Qu[2*]
[1,2]Faculty of Engineering and Technology, Panyapiwat Institute of Management, Nonthaburi, Thailand
E-mail: 6672100041@stu.pim.ac.th, jianqu@pim.ac.th*

Received: September 2024 / Revised: June 12, 2026 / Accepted: June 16, 2026

***Abstract*—This review paper examines explainable image captioning techniques for autonomous driving. It focuses on natural language explanations to improve transparency and explainability. We review more than 20 studies published from 2015 to 2024. These include foundational models, attention mechanisms, Transformer-based methods, and integrations with large language models. The review traces the progress from basic textual descriptions to chain-based reasoning and multimodal fusion. Because of resource limits and the need for low-cost experimentation, we chose the open-source DonkeyCar platform to replicate models. We replicated the ADAPT (Action-aware Driving Caption Transformer) model and compared it with prior methods using standard metrics: BLEU-4, CIDEr, and METEOR. The results show that ADAPT outperforms previous state-of-the-art methods, with CIDEr gains of 31.7 in narration and 33.1 in reasoning. We also successfully reproduced lane-following and obstacle-avoidance tasks on DonkeyCar. Multimodality supports more natural learning for future AI models, but accuracy is limited in most scenarios.**

***Index Terms*—Image Captioning, Explainable Autonomous Driving, Deep Learning, DonkeyCar**

I. Introduction

With the rapid development of computers and the Internet, Artificial Intelligence (AI) has become a crucial part of human life. From deep mathematical and algorithmic research in the 1950s and 1960s, through the development of complex systems in the 1990s, to extensive applications in autonomous driving today, AI has moved from theoretical exploration to practical use. In autonomous driving, AI plays a key role in transforming how we understand transportation, enhancing safety, and increasing convenience [1].

The widespread adoption of autonomous vehicles has led to an urgent need for interpretability. This need arises from several factors. First, autonomous vehicles' role in road accidents presents a fundamental practical issue. Negligence and hazardous driving in traditional vehicles directly affect the safety of passengers and bystanders, making transportation system safety crucial. Furthermore, understanding the motives behind decisions is a fundamental human psychological need. Users are likely to avoid using a system if they cannot understand or trust it. Research shows that systems that are both understandable and perceptible significantly increase user confidence. Frequent system malfunctions without reliable explanations can severely undermine user and public trust. Restoring trust once it is damaged is extremely challenging. Therefore, users need to understand the key decisions autonomous vehicles make in specific situations to build confidence in their decision-making. Confidence in a vehicle's smart decisions will further bolster transparency.

Transparency not only enhances explainability but also ensures the system's decisions and actions comply with current regulations and standards. This is crucial for ensuring fairness, ethical scrutiny, and rational justification of the system's key decisions. These interconnections are essential for the widespread adoption of autonomous vehicles [2].

Specifically, interpretability in autonomous driving systems must account for users' backgrounds and knowledge levels. For example, a user unfamiliar with autonomous driving technology might need a simple, clear explanation of the system's decisions and outcomes to feel satisfied. In contrast, autonomous systems engineers need more detail to understand the vehicle's current functionality and when to activate existing systems. Therefore, analyzing the audience's domain expertise and knowledge traits is crucial to providing appropriate, knowledgeable, and comprehensible explanations.

Recent research and practice have also advanced this field. For instance, "Explaining Explanations: An Overview of Interpretability of Machine Learning" (2019) explores model interpretability and introduces techniques such as Local Interpretable Model-agnostic Explanations (LIME) and integrated gradients, which help enhance the interpretability of autonomous driving systems. Additionally, "Towards Transparent

and Explainable AI: From Interpretability to Explainability" (2021) further expands understanding of AI transparency by emphasizing the relationship between model interpretability and explainability. These findings provide important references for interpreting autonomous driving systems, making their decision-making processes more transparent and understandable [3].

In summary, as autonomous driving technology evolves, the demand for system interpretability grows. By continuously improving explanation techniques and considering user backgrounds and needs, researchers can enhance the transparency and explainability of autonomous driving systems, facilitating broader acceptance in practical applications.

II. Related Studies

A. The Development of Explaining in Autonomous Driving

In recent years, autonomous driving research has increasingly focused on system interpretability and transparency in decision-making processes [4]. From early visual scene descriptions to chain-based reasoning and large language models for natural language explanations, research has deepened, driving the evolution of explainable autonomous driving systems. Below is a chronological review of representative papers in this field, with a detailed analysis of how each study builds on or improves its predecessors [5]. First, "Textual Explanations for Self-Driving Vehicles" is an early study on explainability in autonomous driving. This paper introduced generating natural language explanations for autonomous driving behaviors through visual scene understanding. The system provided a simple yet effective framework that generated coherent textual explanations from perception to decision-making, allowing external observers to clearly understand the vehicle's decision-making process. Its main contribution was establishing a model that converts visual information into natural language, laying the foundation for future studies on generating explanations through language [6].

Building on this, "Reason2Drive: Towards Interpretable and Chain-based Reasoning for Autonomous Driving" extended and improved the previous work [7]. This research introduced a chain-based reasoning model that constructs causal chains tracing each step from sensor data to the final driving behavior. Compared with "Textual Explanations for Self-Driving Vehicles," Reason2Drive not only generated natural-language descriptions of driving behaviors but also emphasized the causal relationships and logical reasoning behind each decision. This significantly enhanced the system's transparency, allowing systematic tracing of the entire decision-making process and increasing trust in autonomous driving systems.

Next, "Advancing Explainable Autonomous Vehicle Systems: A Comprehensive Review and Research Roadmap" offers a thorough examination of explainability technologies in autonomous driving [8]. This paper systematically summarizes current technological developments, focusing on how multimodal fusion and behavior prediction can improve the transparency of autonomous driving systems. Unlike earlier model-specific research, this paper took a macro-level view of the explainability problem, highlighting the strengths and weaknesses of existing methods while outlining challenges and future directions in complex driving scenarios. It provided theoretical guidance for subsequent research and emphasized the importance of multimodal data fusion in enhancing system interpretability.

As technology advanced, "ADAPT: Action-aware Driving Caption Transformer" introduced the Transformer model and significantly improved the accuracy and contextual relevance of driving scene descriptions [9]. ADAPT proposed an "action-aware driving caption transformer" that combined visual input with driving behavior actions to generate natural language explanations related to driving scenes. Compared with earlier methods, ADAPT not only described driving behaviors but also effectively integrated dynamic scene information, making the generated explanations more context-aware and reflective of real-world driving scenarios. The innovation lies in using a Transformer model to enhance contextual understanding and add depth and detail.

Building on this, "DRIVEGPT4: Interpretable End-to-End Autonomous Driving via Large Language Model" uses Large Language Models (LLMs) such as GPT-4 to create even more natural and highly interactive explanations [10]. By incorporating GPT-4, this system could generate more complex natural-language explanations that cover subtle changes in driving scenes. Compared to ADAPT, DRIVEGPT4 emphasized flexibility and adaptability in explanations, especially in complex driving environments, where LLMs gave the system more powerful language-generation capabilities. It not only generated descriptions of driving behaviors but also dynamically explained driving logic across scenarios, demonstrating LLMs' vast potential in autonomous driving.

Finally, "Driving with LLMs: Fusing Object-Level Vector Modality for Explainable Autonomous Driving" further improved explanation precision and detail compared with DRIVEGPT4 [11]. This study fused object-level vector modalities with large language models to generate more precise explanations of driving scenes. Specifically, the system could not only produce global driving-behavior explanations but also provide detailed descriptions of each object

in the driving environment. This object-level explanatory ability significantly improved the system's understanding of complex scenes and offered more actionable insights. Compared with DRIVEGPT4's global scene explanations, "Driving with LLMs" focused more on multimodal fusion, particularly combining visual perception with language generation, and achieved new breakthroughs.

Research on explainability in autonomous driving systems has evolved from simple visual-to-text generation models to integrating reasoning chains and multimodal data. The introduction of Transformer models and large language models has significantly enhanced the system's ability to provide explanations. In particular, LLMs have greatly improved the flexibility and adaptability of explanations, enabling the system to handle complex, dynamic driving scenarios. Future work will continue to improve the accuracy and efficiency of explainability, especially in object-level explanation generation [12].

This review traces the evolution of explainable autonomous driving and compares core transparency techniques. Early methods (e.g., Textual Explanations) provided basic visual-to-text descriptions but lacked explanatory depth. Reason2Drive improved this with chain-based reasoning for better logical traceability, though it was less flexible. ADAPT advanced further by using action-aware Transformers for context-rich captions, outperforming prior models in relevance but requiring more computation. DriveGPT4 introduced LLMs for adaptive, nuanced explanations, addressing earlier rigidity in complex scenes. LLM-driven driving added object-level multimodal fusion, achieving higher precision. Overall, the field has progressed from static text to dynamic, multimodal reasoning with steady gains in interpretability; however, challenges in scalability, real-time performance, and resource efficiency persist for low-cost platforms. Our earlier study on explainable captioning for traffic sign recognition confirms this trend, showing that resource-limited hardware can produce readable justifications for individual road objects [12].

B. The Development of Image Captioning

In recent years, image captioning technology has garnered significant attention and developed rapidly. With the introduction of deep learning models, researchers have continuously optimized image-to-text generation. Below is a review of several representative image captioning papers, presented in chronological order, focusing on their key contributions and developments.

First, "Show and Tell: A Neural Image Caption Generator" is a foundational work in image captioning. Vinyals et al. combine Convolutional Neural Networks (CNNs) for image feature extraction with Recurrent Neural Networks (RNNs) to generate textual descriptions, achieving end-to-end image captioning. This model integrates image representation with a language model, significantly improving the coherence and semantic relevance of generated captions. The innovation lies in merging CNN and RNN and training the model using maximum likelihood estimation. This work laid a solid foundation for subsequent deep learning-based image captioning techniques [13].

Next, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention" builds upon the previous work by introducing the attention mechanism. Xu et al. proposed a model that dynamically focuses on different image regions as it generates each word, producing more precise captions that match specific image features. Compared with "Show and Tell," attention greatly improves the model's ability to handle complex scenes, allowing it to capture image details more effectively. This innovation introduced new ideas to image captioning, particularly excelling at generating descriptions in multi-object scenes [14].

Subsequently, "Self-critical Sequence Training for Image Captioning" introduced self-critical training for sequence-generation tasks. Rennie et al. addressed image caption evaluation by proposing reinforcement learning optimization using metrics like BLEU and CIDEr. This training method lets the model focus more on caption quality rather than relying solely on maximum likelihood estimation. Unlike previous training methods, reinforcement learning improves performance during evaluation, especially for long sequences. This advancement addresses earlier models' tendency to get stuck in local optima during generation [15].

Following this, "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering" introduced bottom-up and top-down attention mechanisms, further improving the model's ability to capture complex information from images [16]. Anderson et al. proposed this dual attention mechanism, which considers global image features and extracts local details, leading to more accurate, content-rich captions. Compared with "Show, Attend, and Tell," this model better balances global information and local features, making the generated captions more aligned with human visual perception.

Building on these advances, "Meshed-Memory Transformer for Image Captioning" introduced a Transformer-based image captioning model. Traditional RNNs struggle with training efficiency and long dependencies, but the Transformer model, with its fully attention-based architecture, addresses these issues. Cornia et al. showed that Transformers can process sequence data in parallel while capturing long-range dependencies between images and text, improving both efficiency and accuracy. This study showed that Transformer-based models

handle complex image information better and scale more effectively [17].

Finally, "Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks" combined object semantic information with visual features and proposed a new pre-training method. Li et al. jointly pre-trained visual and language information, enabling the model to better understand objects' attributes and relationships in images. This approach effectively addressed the issue of separately processing semantic and visual information in traditional methods. Compared with earlier work, Oscar achieved better semantic consistency, improving both the accuracy and diversity of image caption generation [18].

Image captioning technology has evolved from early CNN-RNN models to attention mechanisms and, more recently, to Transformer models and object-semantics pre-training methods. Each step has improved these models' precision, coherence, and efficiency. Future research may focus on better integrating visual and language information, enhancing caption generation in complex scenes, and expanding the use of pre-training methods in multimodal tasks.

C. DonkeyCar

DonkeyCar is an open-source platform designed for small-scale autonomous vehicle experiments, valued for its low cost and flexibility in autonomous driving research. The 2018 paper "Training Neural Networks to Pilot Autonomous Vehicles" showed how Convolutional Neural Networks (CNNs) trained on camera data could control steering and acceleration in DonkeyCar, demonstrating deep learning's potential in small cars.

DonkeyCar's affordability makes it ideal for researchers, educators, and enthusiasts to explore autonomous driving without the expense of commercial platforms. Its open-source design allows for customization and adaptation to diverse research needs.

Building on this, the 2019 paper "End-to-End Deep Learning for Autonomous Driving" proposed an advanced model incorporating Generative Adversarial Networks (GANs) to improve the car's performance in varied environments. Additionally, the 2020 paper "Real-Time Deep Learning for Autonomous Vehicles" optimized real-time performance through edge computing to reduce latency and improve image processing [19].

In 2021, "Multi-Modal Sensor Fusion for Autonomous Vehicles" integrated data from multiple sensors, such as cameras and LiDAR, to improve environmental perception and offered new insights into multi-modal data processing on DonkeyCar [20]. The platform's flexibility enables rapid iteration and testing, making it an excellent tool for developing and validating algorithms in controlled environments. As research progresses, DonkeyCar is likely to support more advanced sensor fusion and real-time processing techniques, expanding its use in education and research.

D. Multi-Modal Technology

Multi-modal technology, which combines image recognition and Natural Language Processing (NLP), has gradually become a key area in artificial intelligence. By integrating visual and linguistic modalities, systems can better understand and generate explanatory content.

Early research primarily focused on image captioning tasks. The 2015 paper "Show, Attend and Tell" combined Convolutional Neural Networks (CNNs) with attention mechanisms to generate natural language descriptions of images, laying the foundation for integrating image and language modalities.

Later, researchers focused on cross-modal learning. The 2018 paper "Bottom-Up and Top-Down Attention" used a dual attention mechanism to link image recognition with language generation, enabling more accurate processing of visual details and improving image captioning and visual question-answering tasks [16].

Generative models further advanced multimodal technology. The 2019 paper "Image Captioning: Transforming Objects into Words" was among the first to apply Transformers to image captioning, showing the power of combining visual and linguistic modalities to produce richer text generation [21].

Recent studies have focused on unified multimodal frameworks. The 2020 paper "Unified Vision-Language Pre-Training for Image Captioning and Visual Question Answering" proposed a unified pre-training model that integrates image captioning and visual question answering, enhancing cross-task performance [22].

Looking ahead, multi-modal technology will increasingly focus on real-time capabilities and generation quality, driving deeper integration of visual and language modalities in practical applications. Recent work on matching traffic sign images with textual captions illustrates this direction on a concrete driving task [23].

III. Research Methodology

We searched Google Scholar, IEEE Xplore, arXiv, and Scopus using keywords such as "explainable autonomous driving," "image captioning autonomous," "multimodal autonomous driving," and "DonkeyCar." We searched for papers published between 2015 and 2024. We included only peer-reviewed papers or high-impact preprints focused on explainability in autonomous driving and image-to-text generation. After screening about 35 papers, we selected 22 core

studies: 10 on autonomous driving explainability, 8 on image captioning, and 4 on multimodal methods and DonkeyCar.

In the experimental part, we replicated the ADAPT model using the pre-processed BDD-X dataset. The setup included Python 3.8, PyTorch, and NVIDIA Apex, with pretrained weights downloaded from Google Drive. We trained for 40 epochs using the default hyperparameters. We evaluated performance on narration and reasoning tasks using BLEU-4, CIDEr, and METEOR metrics. In addition, we successfully reproduced the lane-following and obstacle-avoidance tasks on the DonkeyCar platform. This lays the foundation for integrating explainable image captioning into the DonkeyCar end-to-end autonomous driving system.

IV. Experiments and Results

We replicated the ADAPT model by strictly following the official GitHub repository instructions. ADAPT takes a short driving video clip as input and generates two sentences: A narration describing the vehicle's current action and a reasoning sentence explaining why the action is taken. Fig. 1 shows two examples of this output on BDD-X video clips. In the first example, the model describes the behavior as "the car is turning left" and gives the reason "because the car is making a left turn onto a different street." In the second example, it describes the behavior as "the car is slowing down and coming to a stop" and explains it with "because there is a stop sign."

We began by cloning the repository to obtain the source code and project structure. Next, we created a Conda virtual environment with Python 3.8, installed PyTorch with CUDA support, and then installed NVIDIA Apex to accelerate mixed-precision training. We then installed additional dependencies, including mpi4py and OpenMPI for distributed processing, along with the remaining packages listed in the requirements file. We also prepared a Docker environment, as shown in Fig. 2. We installed Docker, pulled the prebuilt ADAPT image (jxbbb/adapt: Latest), and launched the container with the launch_container.sh script provided in the repository.

We created the necessary folders for checkpoints and models. We downloaded the pretrained weights from the provided Google Drive link and placed them in the checkpoints folder. We also obtained the base Video-Swin Transformer models and caption evaluation tools from their respective Google Drive links and organized them into the appropriate directories. Fig. 3 shows the resulting ADAPT-main/checkpoints/basemodel directory, which contains a checkpoints subfolder for the pretrained weights and a log subfolder for the training logs. Fig. 4 shows the ADAPT-main/models directory, which holds the basemodel, captioning, and Video Swin Transformer subfolders. Fig. 5 shows the ADAPT-main/src/evalcap directory, which contains the caption evaluation tools, including the cider and coco_caption packages and the caption_eval.py and evaluate.py scripts used to compute the metrics.

For the dataset, we downloaded and extracted the preprocessed BDD-X dataset from the provided Baidu link. We then arranged it into the datasets/BDDX/ directory with the required file structure, including frame data and captions. Finally, we split the data into train, validation, and test sets.

We trained for 40 epochs using the multitask script with default hyperparameters for both the Driving Caption Generation and Control Signal Prediction tasks. Finally, we evaluated the model on the BDD-X test split using the provided test script, computing standard metrics BLEU-4, CIDEr, and METEOR separately for the narration (action description) and reasoning (explanation of causes) tasks.

Table I compares state-of-the-art methods. Results adapted from the original paper and confirmed in our replication show that ADAPT achieved superior performance, with CIDEr improvements of 31.7 on narration and 33.1 on reasoning over previous best baselines. This successful replication validates the model's functionality and reproducibility, making it a strong baseline for action-aware driving captioning and explainable autonomous driving research.



Fig. 1. The effect of forming the description

```
#step1: 一键安装docker
#参考: https://blog.csdn.net/qq_27865227/article/details/127246251
wget http://fishros.com/install -O fishros && . fishros

#step2:拉取docker镜像
docker pull jxbbb/adapt:latest

#step3:运行docker镜像（即readme中的命令）
sh launch_container.sh
```

Fig. 2. Prepare the environment for Docker

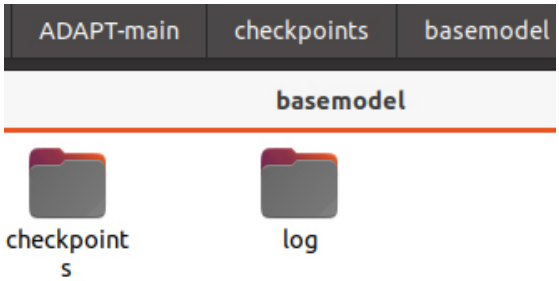


Fig. 3. Prepare the ADAPT-main/checkpoints file and copy the file directory as shown in the figure

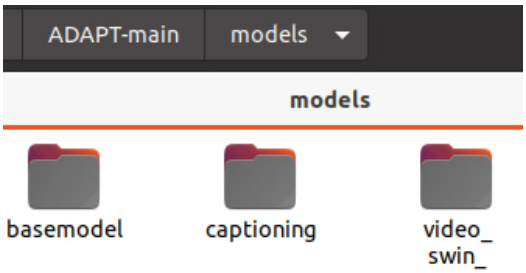


Fig. 4. Prepare the ADAPT-main/models file and copy the file directory as shown in the figure

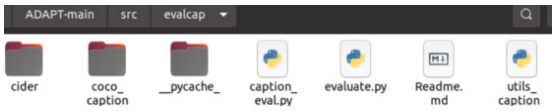


Fig. 5. Prepare the ADAPT-main/src/evalcap file and copy the file directory as shown in the figure

This reproduction of the ADAPT model validates its functionality and demonstrates its capability in video understanding and text generation tasks. The process provides a solid foundation for further model optimization and research into multimodal relationships between video and language, supporting future work in video content understanding.

Compared with existing advanced methods, ADAPT achieves SOTA results in machine evaluation. For example, on the CIDEr indicator, ADAPT is 31.7 higher than previous state-of-the-art methods for action description and 33.1 higher for cause explanation, as shown in Table I.

However, the accuracy is still not high enough for real-world use.

Multimodal fusion makes human-machine interaction more natural, but it needs too much computing power for low-resource devices.

TABLE I
COMPARISON OF RESULTS WITH STATE-OF-THE-ART

Method	Narration			Reasoning		
	BLEU4	CIDEr	METEOR	BLEU4	CIDEr	METEOR
S2VT	30.2	179.8	27.5	6.3	53.4	11.2
S2VT++	27.1	157.0	26.4	5.8	52.7	10.9
SAA	31.8	214.8	29.1	7.1	66.1	12.2
WAA	32.3	215.8	29.1	7.3	69.5	12.2
ADAPT	34.6	247.5	30.6	11.4	102.6	15.0

In the future, we plan to integrate explainable captioning directly into the DonkeyCar platform.

V. Conclusion

Our research systematically reviews explainable image captioning for autonomous driving. First, we explore the necessity of explainability in autonomous driving. Next, we introduce the tasks of explainable autonomous driving and image captioning. After establishing the need for explainability in autonomous driving, we chose the DonkeyCar platform to collect experimental data under our conditions. We also investigated image feature extraction methods, modalities, and datasets. Existing research indicates that multimodality can fully achieve free-form natural language explanations for autonomous driving. Implementing natural language explanations for autonomous driving also suggests greater flexibility in human-machine interaction. Additionally, we validated currently well-performing models through experimental replication. The results show that although related models have improved, their accuracy still does not reach a high level. Therefore, natural language-based explainability for autonomous driving still requires more time and effort. Meanwhile, existing research remains incomplete.

Author Contributions

Conceptualization, N. W. and J. Q.; Methodology, N. W. and J. Q.; Software, N.W. and J. Q.; Validation, N. W. and J. Q.; Formal Analysis, N. W. and J. Q.; Investigation, N. W. and J. Q.; Data Curation, N. W. and J. Q.; Writing—Original Draft Preparation, N.W. and J. Q.; Writing—Review and Editing, N. W. and J. Q.; Visualization, N. W. and J. Q.; Supervision, J. Q.; All authors have read and agreed to the published version of the manuscript. J. Q. is the corresponding Author. The first Author received scholarship support from CPALL for conducting this research in PIM.

References

[1] M. Bojarski et al., "End to end learning for self-driving cars," *arXiv preprint arXiv:1604.07316*, 2016, https://doi.org/10.48550/arXiv.1604.07316

[2] F. Nesti, G. Roossolini, S. Nair, A. Biondi, and G. Buttazzo, "Evaluating the robustness of semantic segmentation for autonomous driving against real-world adversarial patch attacks," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2022, pp. 2280-2289.

[3] S. Ding and J. Qu, "Research on multi-tasking smart cars based on autonomous driving systems," *SN Comput. Sci.*, vol. 4, no. 3, p. 292, Mar. 2023, https://doi.org/10.1007/s42979-023-01740-1

[4] F. Zheng and J. Qu, "TIDCB: Text image dangerous-scene convolutional baseline," *ECTI Trans. Comput. Inf. Technol. (ECTI-CIT)*, vol. 18, no. 3, pp. 280-294, May 2024.

[5] Y. Li and J. Qu, "A novel neural network architecture and cross-model transfer learning for multi-task autonomous driving," *Data Technol. Appl.*, vol. 58, no. 4, pp. 579-598, Apr. 2024.

[6] J. Kim, A. Rohrbach, T. Darrell, J. Canny, and Z. Akata, "Textual explanations for self-driving vehicles," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 577-593.

[7] M. Nie et al., "Reason2Drive: Towards interpretable and chain-based reasoning for autonomous driving," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2024, pp. 292-308.

[8] S. Tekkesinoglu, A. Habibovic, and L. Kunze, "Advancing explainable autonomous vehicle systems: A comprehensive review and research roadmap," *ACM Trans. Human-Robot Interact.*, vol. 14, no. 3, pp. 1-46, Mar. 2025, https://arxiv.org/abs/2404.00019

[9] B. Jin et al., "ADAPT: Action-aware driving caption transformer," in *Proc. IEEE Int. Conf. Robot. Autom.* (ICRA) 2023, pp. 7554-7561, https://arxiv.org/abs/2302.00673

[10] Z. Xu, E. Xie, Z. Zhao, Y. Guo, and K. -Y. Wong, "DriveGPT4: Interpretable end-to-end autonomous driving via large language model," *IEEE Robot. Autum. Lett.*, vol. 9, no. 10. pp. 8186-8193, Oct. 2024, https://doi.org/10.1109/LRA.2024.3440097

[11] L. Chen et al., "Driving with LLMs: Fusing object-level vector modality for explainable autonomous driving," *arXiv preprint arXiv:2310.01957*, 2023, https://doi.org/10.48550/arXiv.2310.01957

[12] N. Wang and J. Qu, "Explainable image captioning for autonomous driving: A traffic sign recognition task," in *Proc. 10th Int. Conf. Bus. Ind. Res. (ICBIR)*, 2025, pp. 35-40.

[13] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 3156-3164.

[14] K. Xu et al., "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. 32nd Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 2048-2057.

[15] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 1179-1195.

[16] P. Anderson et al., "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 6077-6086.

[17] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 10575-10584.

[18] X. Li et al., "Oscar: Object-semantics aligned pre-training for vision-language tasks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 121-137, https://arxiv.org/abs/2004.06165

[19] M. Bechtel, E. McEllhiney, M. Kim, and H. Yun, "DeepPicar: A low-cost deep neural network-based autonomous car," in *Proc. IEEE 24th Int. Conf. Embedded Real-Time Comput. Syst. Appl. (RTCSA)*, 2018, pp. 11-21, https://arxiv.org/abs/1712.08644

[20] K. Huang, B. Shi, X. Li, X. Li, S. Huang, and Y. Li, "Multi-modal sensor fusion for auto driving perception: A survey," *arXiv preprint arXiv:2202.02703*, 2022, https://doi.org/10.48550/arXiv.2202.02703

[21] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image captioning: Transforming objects into words," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2019, pp. 11137-11147. https://arxiv.org/abs/1906.05963

[22] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. J. Corso, and J. Gao, "Unified vision-language pre-training for image captioning and VQA," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2020, pp. 13041-13049, https://arxiv.org/abs/1909.11059

[23] N. Wang and J. Qu, "Research on deep learning-based methods for matching traffic sign images with textual captions," *ECTI Trans. Comput. Inf. Technol. (ECTI-CIT)*, vol. 19, no. 3, pp. 379-391, Jul. 2025, https://doi.org/10.37936/ecti-cit.2025193.262155



Ning Wang is currently pursuing a dual-degree Ph.D. at SIIT, Thammasat University, Thailand, and the Japan Advanced Institute of Science and Technology, Japan. He received an M.Eng. from the Faculty of Engineering and Technology, Panyapiwat Institute of Management, Thailand, in 2025. His research primarily focuses on end-to-end autonomous driving, with a particular emphasis on collecting data for explainable autonomous driving using the DonkeyCar platform. He is also committed to developing secure and reliable autonomous driving systems using Deep Reinforcement Learning.



Jian Qu is an Assistant Professor at the Faculty of Engineering and Technology, Panyapiwat Institute of Management. He received a Ph.D. in 2013 with an Outstanding Performance award from the School of Information Science, Japan Advanced Institute of Science and Technology, Japan. He earned a B.B. with Summa Cum Laude honors from the Institute of International Studies, Ramkhamhaeng University, Thailand, in 2006, and an M.S.I.T. from the Sirindhorn International Institute of Technology, Thammasat University, Thailand, in 2010. He has been a member of the THAI Super AI house committee since 2020. His research interests are natural language processing, intelligent algorithms, machine learning, machine translation, information retrieval, image processing, autonomous driving, and multi-modal AI.