

PROCEEDINGS

Enhancing Interpretability of End-to-End Autonomous Driving via Concept Bottleneck Models

Yiming Cui¹ and Peng Hang^{1,*}

¹Key Laboratory of Road and Traffic Engineering of the State, Ministry of Education, Tongji University, Shanghai, 201804, China

*Corresponding Author: Peng Hang. Email: hangpeng@tongji.edu.cn

ABSTRACT

In recent years, end-to-end (E2E) autonomous driving models have witnessed rapid development due to their powerful capacity for integrating perception and decision-making. However, the inherent "black-box" nature of deep neural networks results in a significant lack of transparency regarding internal decision-making logic, posing critical challenges for model interpretability and safety validation. To address these issues, this study introduces the Concept Bottleneck Model (CBM) framework—originally developed for image classification—and adapts it to the autonomous driving (AD) domain. By constructing an explicit "raw input-capability concept-decision action" mapping, we aim to enhance model transparency without compromising the original performance of decision action generation. On the data processing front, we extract multi-dimensional scene features based on the large-scale nuScenes dataset to establish a foundational library of capability concepts. We further design a multi-stage labeling paradigm that combines Vision-Language Model (VLM) AI expert annotation with human expert verification, ensuring the efficient acquisition of high-quality capability concept labels. During the model training and validation phases, this research designs and evaluates three distinct training paradigms: Hybrid Loss Joint Training, Near-lossless Joint Training, and Strict Posterior Training. These paradigms are utilized to systematically explore the trade-offs and synergies between predictive performance and interpretability. Experimental results demonstrate that the proposed method effectively maintains the planning and control performance of the E2E model while significantly improving the transparency of the decision-making path. Furthermore, the model exhibits the capability to pinpoint the specific sources of performance degradation, providing a robust technical foundation for the development of trustworthy autonomous driving systems.

KEYWORDS

End-to-end autonomous driving; concept bottleneck model; AI expert annotation; near-lossless training

Acknowledgement: The authors would like to thank the reviewers for their valuable comments and suggestions, and colleagues for helpful discussions.

Funding Statement: This work was supported in part by the National Natural Science Foundation of China (52472451, 52302502).

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.



Copyright © 2026 The Author(s). Published by Tech Science Press.

This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.