Video Retrieval of Autonomous Driving Scenarios Based on Large Models

12310520 Rui Yuhan 12310437 Qiao Shihan

Abstract—Video retrieval in autonomous driving scenarios demands models to comprehend complex visual environments and align them with detailed textual descriptions. While multimodal large language models (MLLMs) have demonstrated potential in vision-language tasks, their application to domain-specific video retrieval remains underexplored. Furthermore, the scarcity of publicly available high-quality annotated video data in the autonomous driving domain makes it challenging for models to accurately learn video content semantics from limited data. Additionally, significant differences exist between autonomous driving videos and those in general video datasets, leading to suboptimal performance of generic large models on autonomous driving datasets.

To address these challenges, we evaluate the performance of two distinct pretrained MLLMs on a carefully curated autonomous driving dataset. This dataset features high-quality captions that have undergone multi-round optimization, combining human annotation expertise with advanced automated labeling techniques, while also serving as a benchmark platform for exploring more efficient annotation strategies. Experimental results demonstrate that our approach effectively retrieves relevant video segments based on textual queries, highlighting the potential of MLLMs to enhance retrieval accuracy and scalability.

Index Terms—Autonomous Driving Video Retrieval, Multimodal Large Language Models (MLLMs), Video Understanding, Automated Annotation Optimization

I. INTRODUCTION

II. TWO KNOWLEDGE DISTILLATION METHODS USED

A. DLDKD

B. TeachCLIP

III. ALTERATIONS ON TEACHCLIP

A. On student model

B. On teacher model

IV. CONCLUSION

Autonomous driving video retrieval presents unique challenges due to the complexity of real-world driving scenarios and the scarcity of high-quality annotated datasets. In this work, we explored the potential of multimodal large language models (MLLMs) to bridge the gap between visual perception and textual understanding in autonomous driving contexts. By leveraging a carefully curated dataset with optimized annotations, combining human expertise and automated refinement using Video-LLama2, we demonstrated that pretrained MLLMs can be effectively adapted for domain-specific video retrieval tasks.

Our experiments with two state-of-the-art MLLMs (Vast and Clip-Vip) revealed that fine-tuning on high-quality annotated data significantly improves retrieval accuracy, validating the importance of robust dataset construction. Additionally, our work establishes a benchmark for future research in autonomous driving video understanding, providing a foundation for developing more interpretable and scalable retrieval systems.

Moving forward, we anticipate that further advancements in MLLMs, combined with more sophisticated annotation pipelines, will enhance the generalization capabilities of autonomous driving systems. Future research could explore dynamic annotation strategies, real-time retrieval optimization, and the integration of multi-sensor data (e.g., LiDAR and radar) to further improve model performance in complex driving environments.

Ultimately, this study highlights the critical role of high-quality data and domain-specific adaptation in advancing autonomous driving technologies, paving the way for safer and more intelligent transportation systems.