Large Language Models for Traffic Surveillance Video Understanding

Author: **Bc. Michal Pyšík**
Supervisor: Ing. Ondřej Klíma, Ph.D.

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2025

Motivation

The rapid increase in surveillance camera deployment generates vast amounts of video data, particularly in traffic monitoring. Manually reviewing this footage to find specific events or information is time-consuming and inefficient. Existing automated methods often focus on predefined event detection and lack the flexibility to handle nuanced, natural language queries or provide deeper semantic understanding of the video content.

Proposed Solution

1. The embedding search is handled by an image-text multimodal embedding model (e.g., CLIP) and a vector database (Milvus).

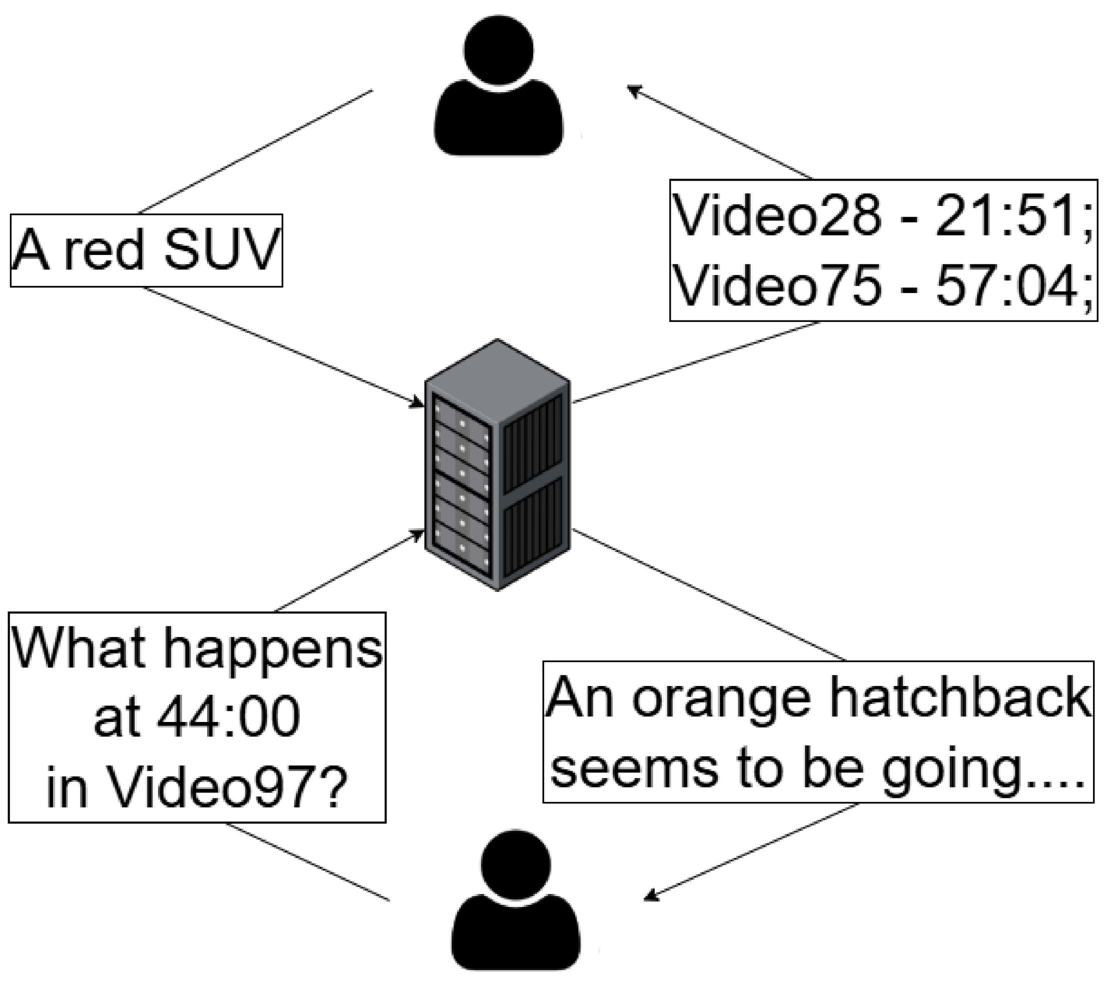

Figure 1: Diagram showing the two different ways a user interacts with the system.

2. Video analysis is handled by a multimodal large language model (MLLM) capable of understanding videos (e.g., GPT-4o), in the form of an interactive chat with the user.

The Implemented System

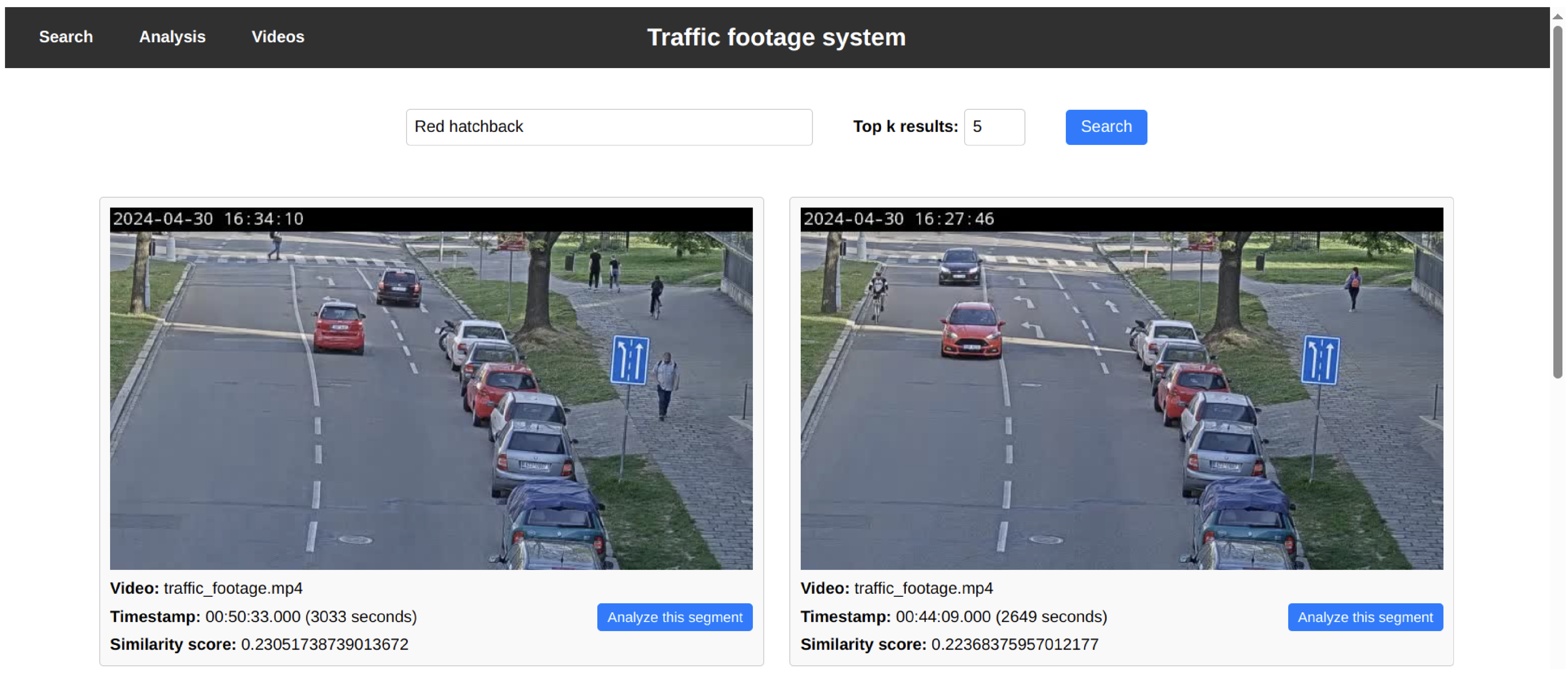

Figure 2: The implemented image-text embedding search (CLIP).

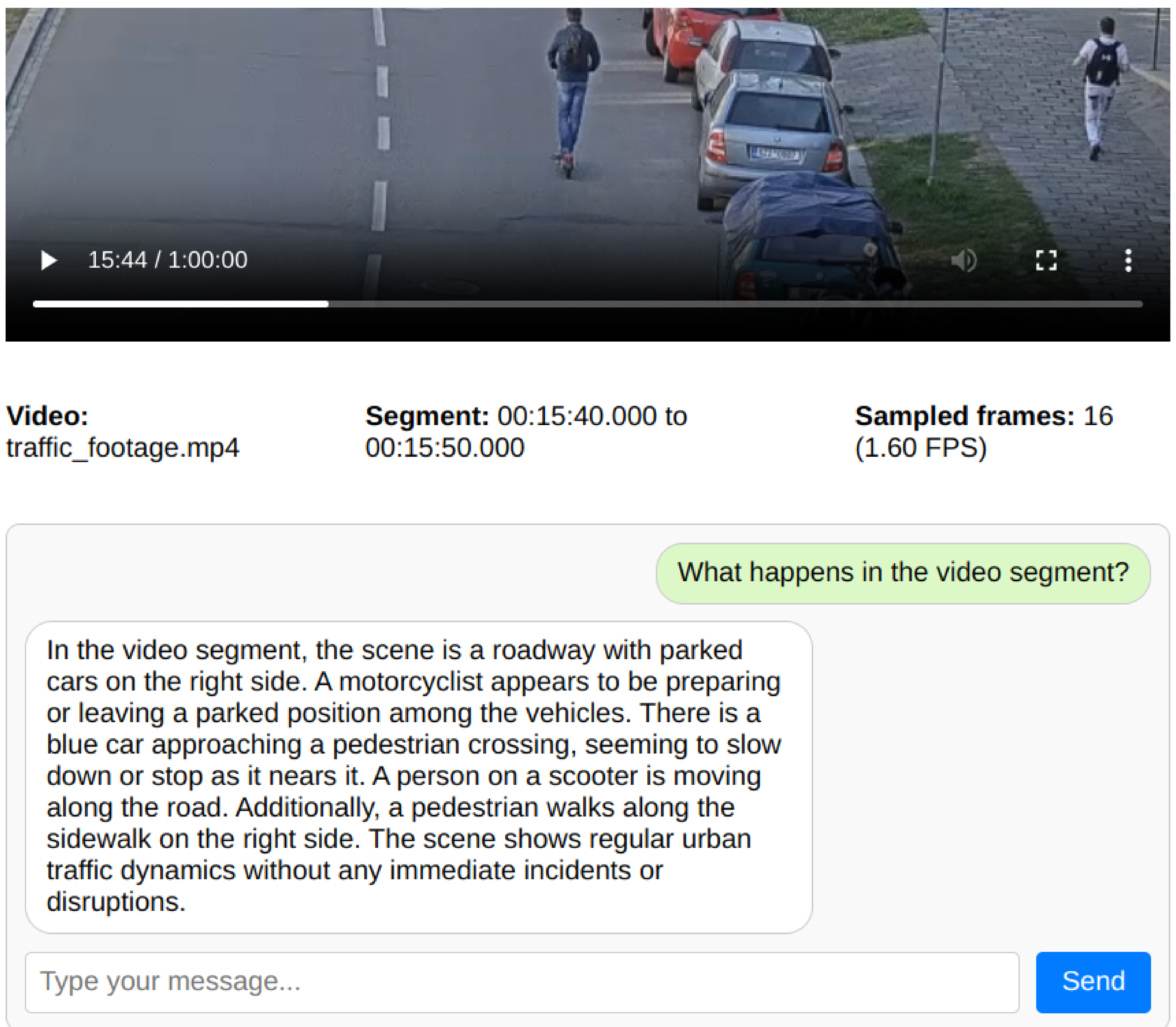

Figure 3: The implemented interactive MLLM video analysis (GPT-4o).

Benchmarks of the Available AI Models

The system allows the user to choose one of the four available multimodal embedding models (CLIP, SigLIP, ALIGN, BLIP) and one of the four available MLLMs (LLaVA-OneVision, GPT-4o, VideoLLaMA 3, Qwen2.5-VL). The performance of these models (image-text retrieval for the multimodal embedding models, video question answering for the MLLMs) in the context of traffic footage was benchmarked on traffic-related datasets (CARS196, Czech traffic signs, SUTD-TrafficQA).

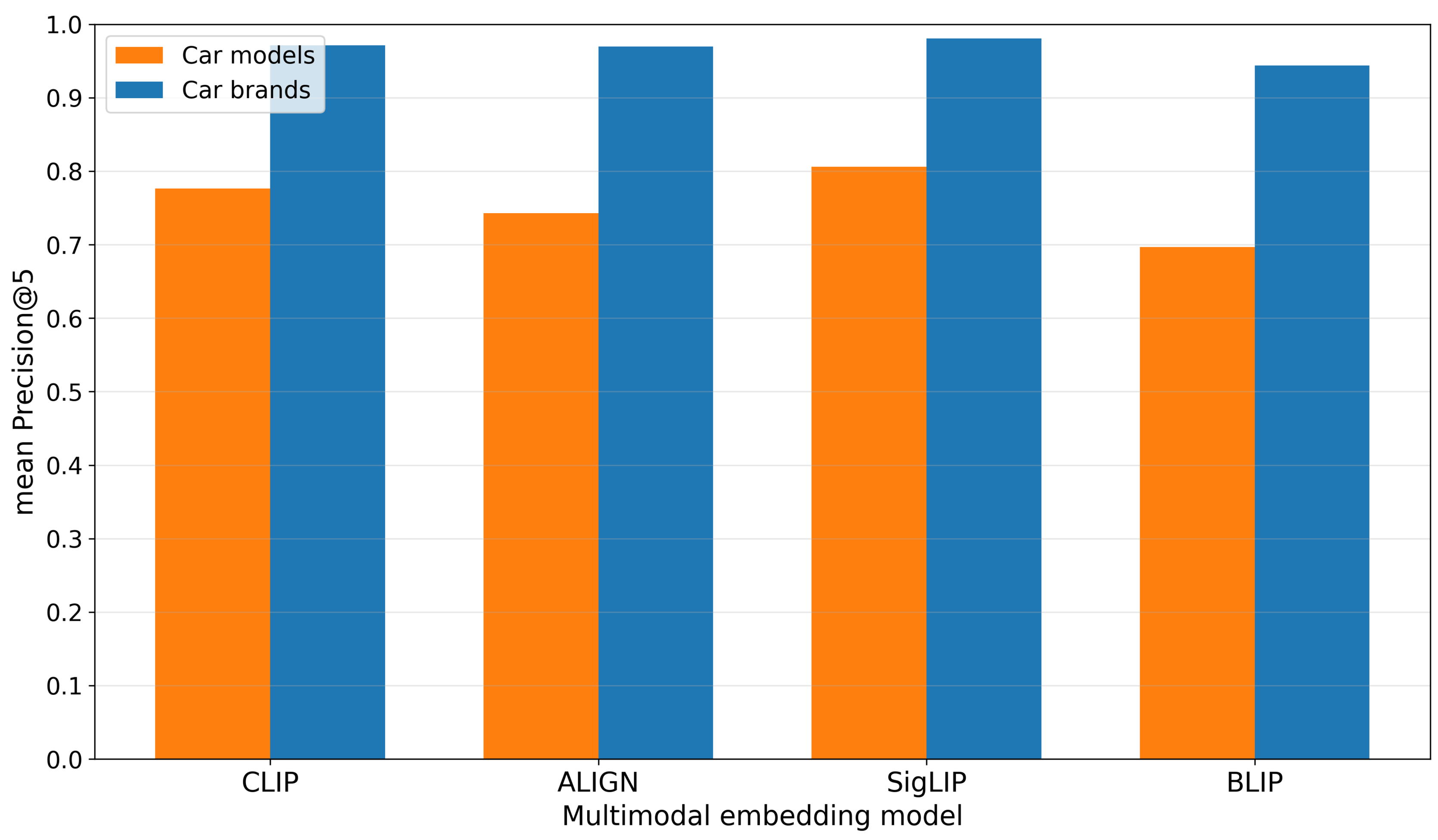

Figure 4: Results of the CARS196 benchmark. The plot shows the mean Precision@5 score for each model, indicating how effectively it retrieves relevant images of cars based on textual queries.

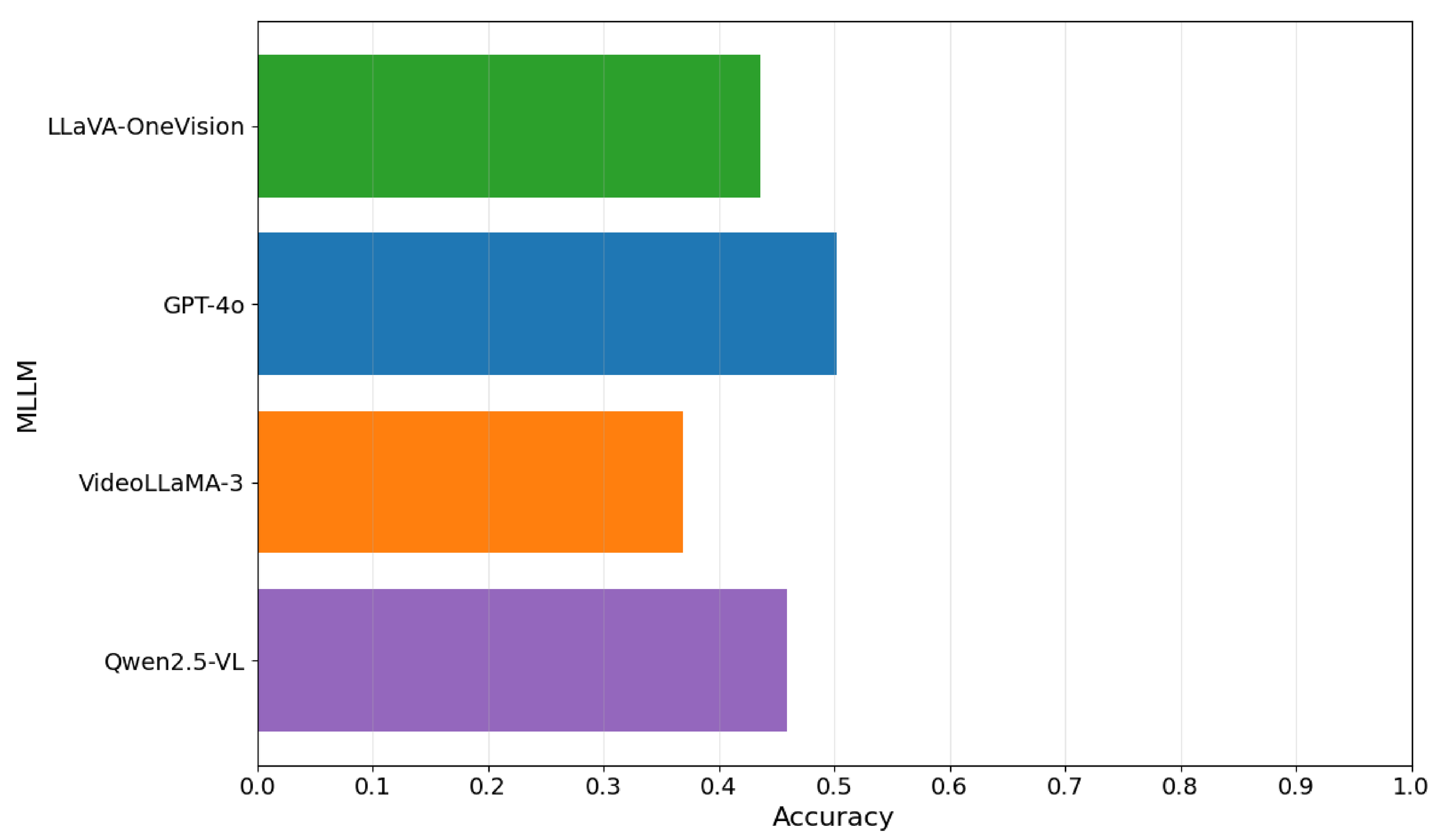

Figure 5: Results of the SUTD-TrafficQA benchmark. The plot shows the accuracy of each model, indicating how well it can understand traffic footage and select the correct answer for single-choice questions.