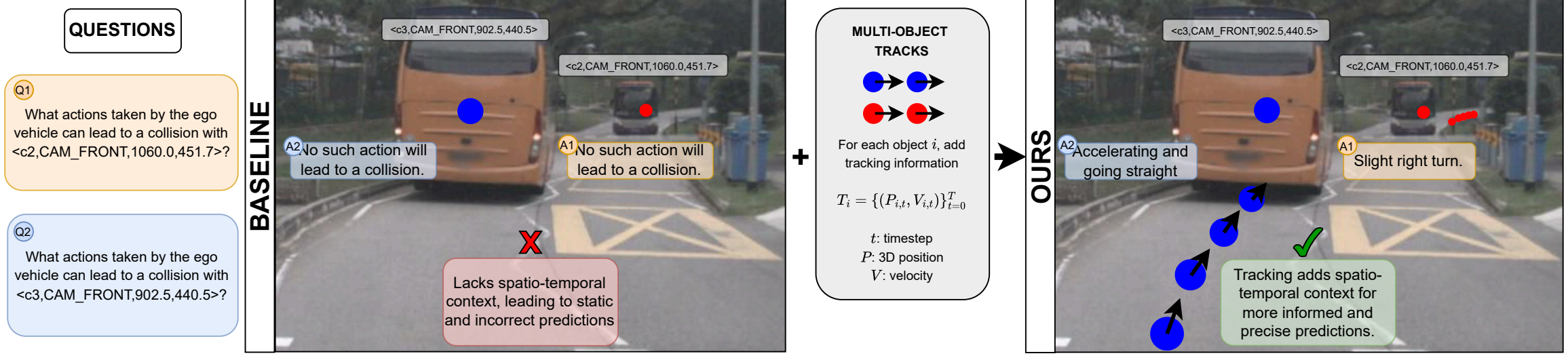
Tracking Meets Large Multimodal Models for Driving Scenario Understanding

Ayesha Ishaq[1], Jean Lahoud[1], Fahad Shahbaz Khan[1,2], Salman Khan[1,3], Hisham Cholakkal[1], Rao Muhammad Anwer[1]

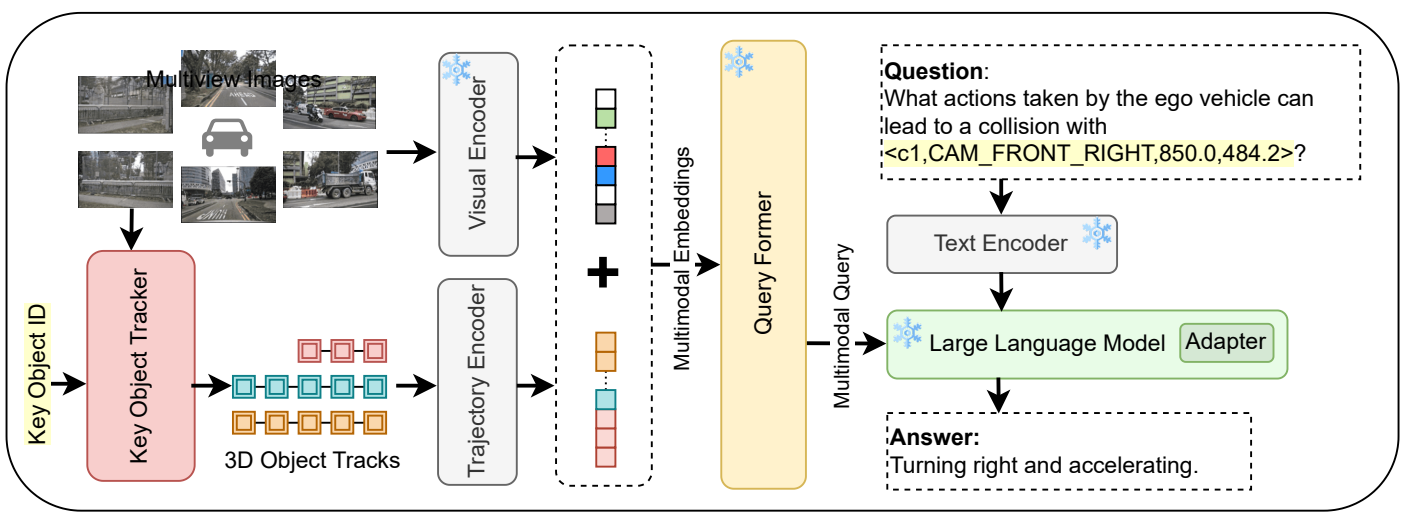
[1]Mohamed Bin Zayed University of Artificial Intelligence,
[2]Linköping University, [3] Australian National University

BMVC 2025

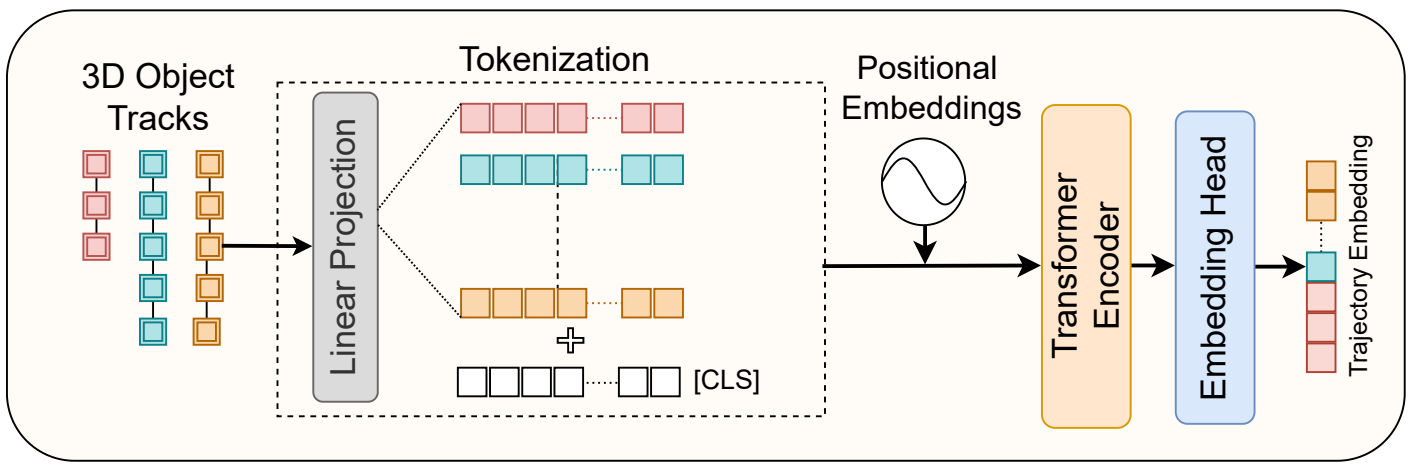
Code, model and dataset available here:



Our system integrates visual, trajectory, and ego-motion information through visual and trajectory encoders. The input consists of multi-view images, which are processed to obtain visual embeddings, while a 3D tracker generates key object tracks. These tracks are then used to extract spatiotemporal embeddings, which are fused with the visual embeddings and transformed through a query former module. These multimodal embeddings are then passed to the large language model enhanced with adapters to align visual, tracking and textual modalities, enabling contextual reasoning and task-specific answers.

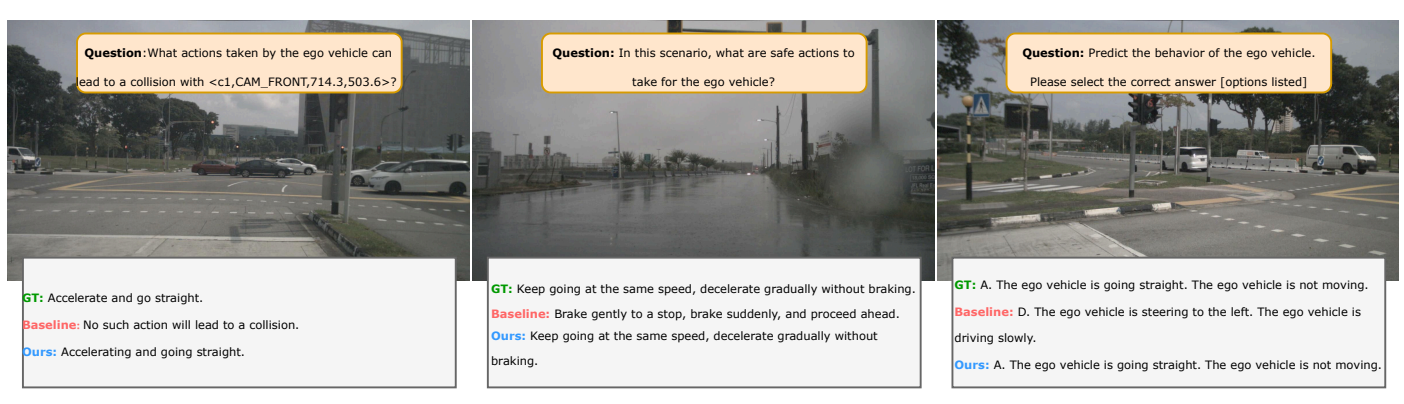


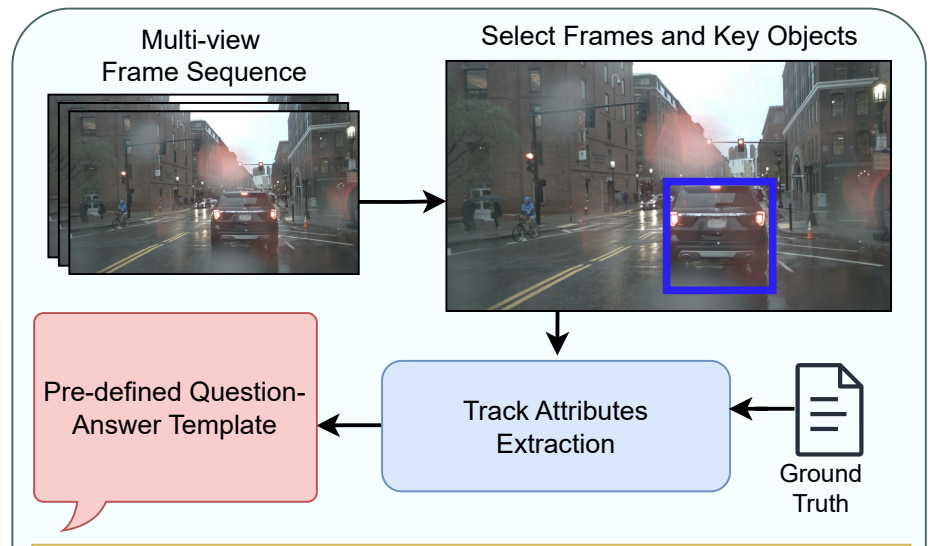
Our system integrates visual, trajectory, and ego-motion information through visual and trajectory encoders. The input consists of multi-view images, which are processed to obtain visual embeddings, while a 3D tracker generates key object tracks. These tracks are then used to extract spatiotemporal embeddings, which are fused with the visual embeddings and transformed through a query former module. These multimodal embeddings are then passed to the large language model enhanced with adapters to align visual, tracking and textual modalities, enabling contextual reasoning and task-specific answers.



	Method	Accuracy	ChatGPT	Bleu 1	Bleu 2	Bleu 3	Bleu 4	ROUGE L	CIDEr	Match	Final Score
Val Set	DriveLM	0.454	55.272	0.680	0.625	0.566	0.512	0.672	0.006	33.750	0.464
	DriveLM + Video	0.667	56.000	0.670	0.606	0.542	0.485	0.676	0.014	39.250	0.519
	Ours	**0.818**	**67.090**	**0.730**	**0.670**	**0.610**	**0.552**	**0.734**	**0.140**	**43.500**	**0.611**
Test Set	DriveLM	0.501	51.396	0.614	0.559	0.506	0.455	0.658	0.033	17.883	0.421
	DriveLM + Video	0.523	55.352	0.706	0.642	0.581	0.522	0.701	0.040	29.074	0.472
	Ours	**0.596**	**58.438**	**0.724**	**0.658**	**0.593**	**0.531**	**0.721**	**0.044**	**35.730**	**0.515**

On the DriveLM-nuScenes test set, our method demonstrates a substantial advantage by integrating 3D tracking information. Compared to the baseline, our approach achieves a 9.5 percentage point gain in accuracy and a 9.4 percentage point increase in the overall final score. This highlights the critical role of spatiotemporal context for understanding complex real-world driving scenarios.





We make use of nuScenes ground truth and multi-view frames to generate tracks and question-answer pairs related to track attributes for pretraining our proposed trajectory encoders.

Method	Accuracy	ChatGPT	Bleu 1	ROUGE L	CIDEr	Final Score
DriveLM	0.908	72.824	0.732	0.712	1.145	0.714
Ours	**0.917**	**77.972**	**0.753**	**0.736**	**1.617**	**0.751**

We also validated our method on the DriveLM-CARLA dataset. The results show our approach adapts to the simulated environment, outperforming the baseline with a 3.7 point gain in the final score.