Task Interference in VLMs for Autonomous Driving: When Better Perception Hurts Planning

Abstract

Vision-language foundation models (VLMs) show strong potential in autonomous driving for scene understanding and decision-making, yet their cross-task performance remains inconsistent. This work presents a systematic study of task interference in VLMs for autonomous driving, revealing a key trade-off: fine-tuning for better perception often degrades planning accuracy. We introduce an evaluation framework that decouples perception and planning to measure interference precisely. Using a multi-source question-answering dataset from diverse driving datasets, we fine-tune state-of-the-art VLMs on action descriptions. While fine-tuned models improve decision explanation quality, they exhibit measurable declines in planning compared to zero-shot counterparts. Experiments across multiple architectures confirm this perception–planning trade-off as a general phenomenon driven by attention conflicts and representation divergence. Our findings provide the first empirical validation of foundation model interference in autonomous driving and highlight critical implications for reliable deployment in safety-critical environments.

1 Introduction

The advent of vision-language foundation models has ushered in a new era of artificial intelligence capabilities, with models like GPT-4V (OpenAI 2023), Gemini (Team et al. 2025a), and LLaMA-Vision (Touvron et al. 2023) demonstrating remarkable performance across diverse visual understanding tasks. Recent works have highlighted a concerning limitation of foundation models: “inconsistent performance across tasks” during deployment (Zhou et al. 2024). While these models can handle multiple computer vision and reasoning tasks, they often exhibit unpredictable performance variations when adapted to specific domains or fine-tuned for particular capabilities. This inconsistency becomes particularly problematic in autonomous driving (Hu et al. 2023; Sima et al. 2025), where a single model may need to simultaneously excel at scene understanding, object detection, action recognition, trajectory prediction, and motion planning. The standard approach to adapting foundation models for specific applications involves fine-tuning on domain-specific datasets to improve performance on target tasks. In autonomous driving, this typically means training models on datasets like NuScenes (Caesar et al. 2020), Argoverse (Wilson et al. 2023), or BDD100K (Yu et al. 2020) to enhance their understanding of driving scenarios, traffic rules, and vehicle behaviors. While this approach has shown success in improving performance on specific metrics, the broader implications for multi-task performance remain largely unexplored.

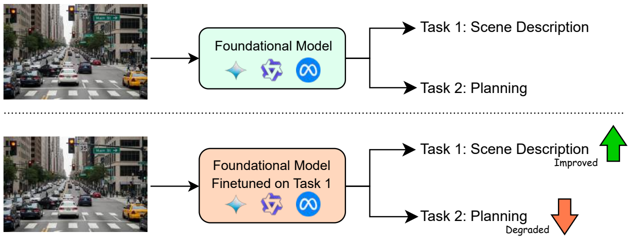


Figure 1: Fine-tuning VLMs to improve scene understanding significantly degrades their planning performance, revealing a fundamental challenge for deploying VLMs

This gap in understanding is particularly concerning given the recent trend toward end-to-end autonomous driving systems, where a single model is responsible for the entire perception-to-action pipeline (Zheng et al. 2024; Xu et al. 2024). These systems promise simplicity and potentially better integration between different components, but they also concentrate all task dependencies within a single model architecture. If fine-tuning for one capability inadvertently degrades performance on another, the consequences could be catastrophic in real-world deployment.

In this work, we address this critical knowledge gap by providing the first systematic empirical analysis of task interference in vision-language models for autonomous driving. Specifically, we investigate the counterintuitive phenomenon where improving a model’s ability to describe and explain driving actions—a seemingly complementary skill to planning—actually degrades its planning performance.

Our contributions are two-fold: First, we develop a systematic evaluation framework that enables precise measurement of task interference effects by decoupling perception and planning capabilities. Second, we provide comprehensive empirical evidence of the perception-planning tradeoff across multiple state-of-the-art vision-language models Qwen (Yang et al. 2025), LLaMA(Touvron et al. 2023), Gemma (Team et al. 2025b)) and diverse driving datasets (NuScenes(Caesar et al. 2020), Indian Driving Dataset (Dokania et al. 2022), Argoverse(Wilson et al. 2023)).

2 Related Work

2.1 Foundation Model Limitations and Task Interference

Recent work has begun to identify fundamental limitations of foundation models beyond their impressive capabilities. Bommasani et al. (Bommasani et al. 2021) highlighted several challenges including inconsistent performance across tasks and domains. More recently, several studies have provided empirical evidence of these limitations in specific contexts (Mukhoti et al. 2024; Liang et al. 2022).

The phenomenon of catastrophic forgetting in neural networks (McCloskey and Cohen 1989; Kirkpatrick et al. 2017) is closely related to task interference in foundation models. When fine-tuning on specific tasks, models may lose previously acquired capabilities, leading to degraded performance on tasks not included in the fine-tuning data. While continual learning approaches have been proposed to address this issue (De Lange et al. 2021; Wang, Zhang, and Zhu 2024), their application to complex multi-modal foundation models in safety-critical domains remains largely unexplored.

2.2 Evaluation Methodologies for Autonomous Driving Models

Evaluating autonomous driving models presents unique challenges due to the multi-faceted nature of driving tasks and the safety-critical requirements (Janai et al. 2020). Traditional metrics like BLEU scores for language tasks or mAP for detection tasks may not capture the complex interdependencies between different capabilities required for safe driving (Caesar et al. 2020).

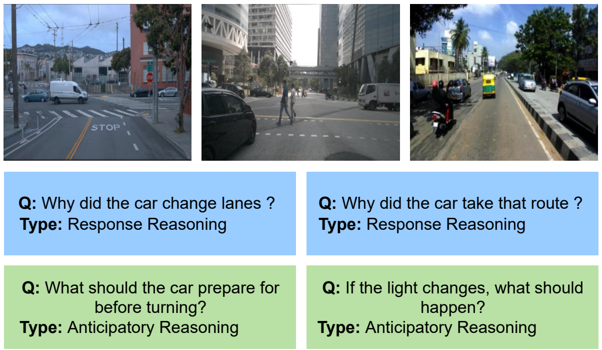
Recent work has proposed more comprehensive evaluation frameworks for autonomous driving systems. DriveLM (Sima et al. 2025) introduced a structured approach to evaluating language-grounded driving capabilities, while other works have focused on closed-loop evaluation in simulation environments (Dosovitskiy et al. 2017; Caesar et al. 2020). However, existing evaluation methodologies do not adequately address the challenge of measuring task interference effects, particularly the trade-offs between complementary capabilities like scene understanding and motion planning.

2.3 Gap Analysis

While previous work has demonstrated the potential of vision-language models for autonomous driving and identified general limitations of foundation models, several critical gaps remain. First, there is a lack of systematic investigation into how fine-tuning for specific driving capabilities affects performance on other tasks within the same model. Second, existing evaluation frameworks do not provide mechanisms to isolate and measure task interference effects. Third, the relationship between perceptual reasoning tasks (such as action description) and planning capabilities has not been empirically studied.

Our work addresses these gaps by providing the first systematic analysis of task interference in vision-language models for autonomous driving, with a specific focus on the perception-planning trade-off. We introduce evaluation methodologies that enable precise measurement of these effects and provide empirical evidence across multiple models and datasets. This work establishes a foundation for understanding and addressing task interference challenges in the deployment of foundation models for safety-critical autonomous driving applications.

3 Methodology



Q: Why did the car change lanes ?
Type: Response Reasoning

Q: Why did the car take that route ?
Type: Response Reasoning

Q: What should the car prepare for before turning?
Type: Anticipatory Reasoning

Q: If the light changes, what should happen?
Type: Anticipatory Reasoning

Figure 2: QuADrive annotation examples showing response reasoning and anticipatory reasoning categories for decision-explaining tasks.

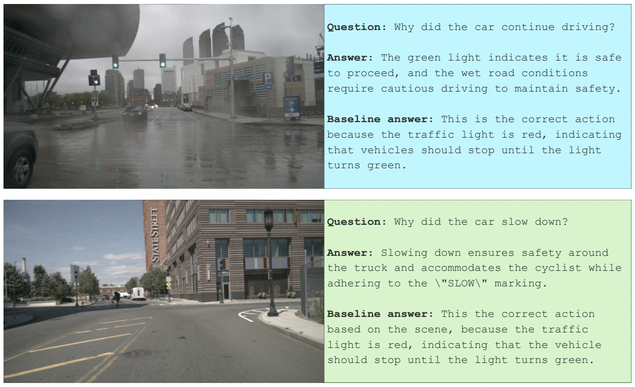
3.1 Dataset Construction and Annotation

We introduce **QuADrive**, a multi-source corpus for training vision-language models in explainable autonomous driving. QuADrive contains 8,346 short driving scenes with decision-explanation question-answer dialogues, aggregated from five datasets: NuScenes, IDD-Temporal, IDD Missing TS, RUGD (Wigness et al. 2019), and Argoverse. This diverse composition covers urban, rural, and adverse-weather settings, supporting robust cross-domain generalization.

Each video is uniformly sampled into ordered frames and presented as a single multi-image user turn in the native chat format of each vision-language model. Using GPT-4o as an annotation oracle, we generate three fields: *Scene Description*, *Action Taken*, and *Justification*. The *Action Taken* is programmatically converted into a question such as “Why did the car slow down?” or “Why did the car remain stationary?”, paired with the *Justification* as the ground-truth explanation.

For cross-dataset evaluation, we use three reasoning benchmarks DriveLM, DriveBench, and QuADrive (ours) each split 90/10 for train-validation. Models are trained on each dataset separately and evaluated across all validation sets, forming nine train-test combinations to analyze within- and cross-dataset performance. Planning evaluation follows the Light-Emma (Qiao et al. 2025) framework on the same splits.

3.2 Fine-tuning Methods and Evaluation



Question: Why did the car continue driving?

Answer: The green light indicates it is safe to proceed, and the wet road conditions require cautious driving to maintain safety.

Baseline answer: This is the correct action because the traffic light is red, indicating that vehicles should stop until the light turns green.

Question: Why did the car slow down?

Answer: Slowing down ensures safety around the truck and accommodates the cyclist while adhering to the \"SLOW\" marking.

Baseline answer: This the correct action based on the scene, because the traffic light is red, indicating that the vehicle should stop until the light turns green.

Figure 3: **Qualitative comparison on the reasoning task before and after fine-tuning.** Baseline answer refers to answers generated by the original model (Qwen2.5-VL-7B), while Answer reflects responses by the model after being fine-tuned using GPT-generated outputs as ground truth. The examples illustrate how fine-tuning enhances the model's ability to reason about visual scenes and justify its decisions.

We evaluate three vision-language model families—Qwen2.5-VL-7B-Instruct, Llama-3.2-11B-Vision-Instruct, and PaliGemma-3-9B—in zero-shot and fine-tuned settings. All models process multi-image inputs via native chat templates, using 4-bit quantization (bitsandbytes) for single-GPU training on an NVIDIA L40S (46GB VRAM). The sequence length is capped at 2,048 tokens with gradient checkpointing. Four fine-tuning strategies are compared: **Full fine-tuning** (all parameters updated via cross-entropy), **LoRA** (Hu et al. 2021) (rank-r adapters with $r = 16$, $\alpha = 16$), **L2-SP** (regularization $\lambda\|\mathbf{w} - \mathbf{w}_0\|_2^2$, $\lambda = 0.01$), and **SAM** (Liu et al. 2025) (LoRA with sharpness-aware minimization, $\rho = 0.05$). All use AdamW (2×10^{-4}), batch size 2, 4-step accumulation, and linear decay. Models are trained on DriveLM, DriveBench, and WhyDrive with 2 warmup epochs and early stopping; Qwen2.5-VL-7B runs 15 epochs, while Llama and Gemma train for 20.

We assess two tasks: **decision explanation** and **trajectory planning**. The former evaluates natural language justifications for driving actions using GPT-4o generated ground truths on NuScenes and cross-domain sets (IDD-Temporal, IDD Missing TS, RUGD, Argoverse). The latter adopts the Light-Emma framework, where models predict six future waypoints at 0.5s intervals as (v, κ) tuples, evaluated via L_2 error at 1s, 2s, and 3s on NuScenes-mini. Planning data are disjoint from QuADrive to isolate task interference. Llama and Gemma follow the original Light-Emma prompt, while Qwen uses a shortened version to reduce violations. Minor format repairs ensure valid numeric outputs for fair L_2 evaluation. Two baselines—zero-shot and decision-only fine-tuning—are compared, with results normalized to zero-shot performance to measure adaptation and interference effects.

4 Results

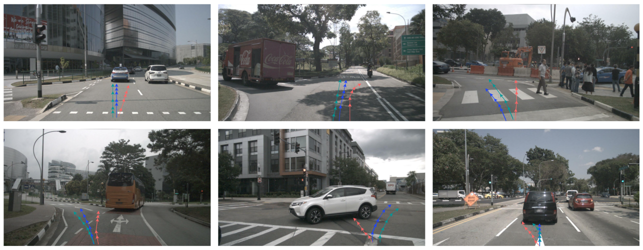


Figure 4: **Qualitative comparison of waypoint predictions across diverse driving scenarios. Ground truth trajectories (blue), zero-shot model predictions (green), and fine-tuned model predictions (red)** of Gemma in LORA way are overlaid on front-camera images. Zero-shot predictions closely track ground truth across all scenarios, while fine-tuning for decision-explaining systematically compromises geometric reasoning and trajectory planning capabilities.

Tables 1 and 2, together with Figure 4, present a comprehensive view of task interference and generalization behavior in vision-language models for autonomous driving. Fine-tuning on QuADrive substantially boosts decision-explaining quality, achieving 0.80–0.89 BERT scores versus 0.65–0.68 for zero-shot baselines—a 23–32% improvement—supported by consistent BLEU and ROUGE gains. However, this comes at the cost of severe planning degradation: response errors surge from 8–15% to over 40–54%, and mean L_2 trajectory error increases by 30–46% across Qwen, Llama, and Gemma. Even advanced regularization (LoRA, L2-SP, SAM (Liu et al. 2025)) provides only partial relief, confirming that task interference is structural rather than optimization-specific.

Cross-domain evaluation further reveals asymmetry in generalization. QuADrive-trained models transfer effectively to DriveLM and DriveBench (0.86–0.87), while models trained on these datasets drop to 0.65–0.71 on QuADrive. QuADrive's diverse, high-quality annotations enable superior generalization, surpassing DriveLM and DriveBench even in-domain. LoRA retains 95–98% of in-domain accuracy versus 92–94% for full fine-tuning, indicating modest robustness under distribution shift. Qualitative trajectory visualizations reinforce these findings: zero-shot models (green) generate smooth, kinematically consistent paths, whereas fine-tuned models (red) exhibit spatial drift, premature termination, or invalid trajectories. Frequently encountered scenarios like pedestrian crossing scene (top-right) in Figure 4 demonstrates output format degradation beyond mere spatial errors. Collectively, these results highlight a persistent perception–planning trade-off—fine-tuning enhances reasoning alignment but undermines geometric and temporal consistency, exposing task interference as a fundamental limitation in multi-task VLM adaptation for autonomous driving.

Train	Test	Qwen-2.5-VL-7B					Llama-3.2-11B					Gemma-2-9B				
		Zero-shot	Full FT	LoRA	L2-SP	SAM	Zero-shot	Full FT	LoRA	L2-SP	SAM	Zero-shot	Full FT	LoRA	L2-SP	SAM
DriveLM	DriveLM	0.65	0.82	0.80	0.79	0.81	0.62	0.79	0.77	0.76	0.78	0.59	0.75	0.73	0.72	0.74
	DriveBench	0.63	0.81	0.79	0.78	0.80	0.60	0.78	0.76	0.75	0.77	0.57	0.74	0.72	0.71	0.73
	QuADrive	0.68	0.89	0.87	0.86	0.88	0.65	0.86	0.84	0.83	0.85	0.62	0.82	0.80	0.79	0.81
DriveBench	DriveLM	0.58	0.71	0.69	0.68	0.70	0.55	0.68	0.66	0.65	0.67	0.52	0.64	0.62	0.61	0.63
	DriveBench	0.56	0.69	0.67	0.66	0.68	0.53	0.66	0.64	0.63	0.65	0.50	0.62	0.60	0.59	0.61
	QuADrive	0.63	0.84	0.82	0.81	0.83	0.60	0.81	0.79	0.78	0.80	0.57	0.77	0.75	0.74	0.76
QuADrive (Ours)	DriveLM	0.68	0.85	0.83	0.82	0.84	0.65	0.82	0.80	0.79	0.81	0.66	0.83	0.81	0.80	0.82
	DriveBench	0.67	0.84	0.82	0.81	0.83	0.64	0.81	0.79	0.78	0.80	0.65	0.82	0.80	0.79	0.81
	QuADrive	**0.72**	**0.93**	**0.91**	**0.90**	**0.92**	**0.69**	**0.90**	**0.88**	**0.87**	**0.89**	**0.70**	**0.94**	**0.92**	**0.91**	**0.93**

Table 1: **Comparison of BERTScore performance for scene description Q&A across three models** (Qwen-2.5-VL-7B, Llama-3.2-11B, Gemma-2-9B) fine-tuned using five methods (Zero-shot, Full FT, LoRA, L2-SP, SAM) on three datasets (DriveLM, DriveBench, QuADrive). Training on the QuADrive (Ours) dataset consistently yields the highest scores across all test sets, outperforming others by 8–15%, indicating superior generalization and action description quality.

Model	Method	NuScenes					Waymo					Avg Degradation from Zero-shot
		Err	L2@1s	L2@2s	L2@3s	L2avg	Err	L2@1s	L2@2s	L2@3s	L2avg	
		(%)	(m)	(m)	(m)	(m)	(%)	(m)	(m)	(m)	(m)	
Qwen-2.5-VL-7B	Zero-shot	8.2	0.91	1.47	2.13	1.50	11.4	0.97	1.54	2.28	1.60	*baseline*
	Full FT	56.8	1.32	2.21	3.52	2.35	64.2	1.48	2.42	3.81	2.57	↓ 48.6%
	LoRA	54.3	1.28	2.15	3.42	2.28	61.7	1.45	2.38	3.72	2.52	↓ 46.1%
	L2-SP	52.1	1.23	2.06	3.25	2.18	59.3	1.39	2.28	3.58	2.42	↓ 43.9%
	SAM	50.7	1.19	1.98	3.18	2.12	57.8	1.35	2.21	3.47	2.34	↓ 42.5%
Llama-3.2-11B	Zero-shot	12.5	0.95	1.52	2.21	1.56	14.8	1.02	1.61	2.35	1.66	*baseline*
	Full FT	42.1	1.38	2.12	3.25	2.25	47.5	1.55	2.35	3.51	2.47	↓ 29.6%
	LoRA	38.7	1.35	2.08	3.18	2.20	43.2	1.52	2.31	3.44	2.42	↓ 26.2%
	L2-SP	36.4	1.29	1.98	3.05	2.11	40.9	1.45	2.21	3.28	2.31	↓ 23.9%
	SAM	35.1	1.25	1.92	2.98	2.05	39.2	1.41	2.15	3.21	2.26	↓ 22.6%
Gemma-2-9B	Zero-shot	15.1	1.03	1.68	2.45	1.72	17.3	1.11	1.75	2.58	1.81	*baseline*
	Full FT	45.6	1.44	2.28	3.58	2.43	52.1	1.61	2.51	3.87	2.66	↓ 30.5%
	LoRA	42.1	1.41	2.23	3.51	2.38	48.6	1.58	2.47	3.78	2.61	↓ 27.0%
	L2-SP	39.8	1.34	2.12	3.35	2.27	45.7	1.51	2.35	3.61	2.49	↓ 24.7%
	SAM	**38.3**	**1.31**	**2.06**	**3.28**	**2.22**	**43.9**	**1.48**	**2.29**	**3.54**	**2.44**	**↓ 23.2%**

Table 2: **Planning Performance Degradation Across Fine-Tuning Methods :** All fine-tuning methods cause substantial planning degradation across all prediction horizons. SAM performs best but still shows 22-42% performance loss. Err = Response error rate; L2@Xs = L2 trajectory error at X seconds; L2avg = average of L2@1s, L2@2s, and L2@3s.

5 Conclusion

We provide a systematic analysis of task interference in vision-language models for autonomous driving, demonstrating that fine-tuning for decision-explaining degrades planning performance across three model families (Qwen, Llama, Gemma) and four adaptation methods (Full FT, LoRA, L2-SP, SAM). Even state-of-the-art preservation techniques fail to prevent catastrophic capability loss, with response errors increasing 6-7x and trajectory accuracy declining 40-50%. This consistency across architectures, scales, and datasets indicates task interference is a fundamental challenge rather than an artifact of specific training procedures. Our mechanistic analysis identifying attention conflicts, representation divergence, and format specialization informs proposed mitigation strategies including multi-task balancing and modular architectures. As autonomous driving companies increasingly adopt foundation models, understanding and addressing these interference effects becomes essential for safe deployment. The QuADrive dataset and evaluation framework we introduce enable rigorous investigation of multi-task learning dynamics in safety-critical applications, establishing empirical evidence.

Our findings underscore a broader challenge for foundation model adaptation: fine-tuning that enhances one capability can systematically suppress others, undermining the very generalization such models were designed to provide. This phenomenon suggests that pre-trained representations encode a delicate balance across tasks—one that is easily disrupted by narrow optimization. Our study focuses on specific task combinations (action description and trajectory prediction) and may not generalize to all possible task interference scenarios. Future work should investigate interference effects across broader sets of autonomous driving tasks, including object detection, traffic sign recognition, and behavior prediction.

References

Bommasani, R.; Hudson, D. A.; Adhikari, E.; Altman, P.; Argyris, K.; Aronzon, E.; Beer, S.; Belle, C.; Berman, J. E.; Blum, A.; et al. 2021. On the Opportunities and Risks of Foundation Models. *arXiv preprint arXiv:2108.07258*.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuScenes: A multimodal dataset for autonomous driving. arXiv:1903.11027.

De Lange, M.; Aljundi, R.; Masana, M.; Parisot, M.; Slabaugh, G.; Tuytelaars, T.; Van Gool, L.; Williams, C. K.; Yang, J.-Q.; and Zhu, Y. 2021. A Continual Learning Survey: Defying Forgetting in Classification Tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7): 3376–3397.

Dokania, S.; Hafez, A. H. A.; Subramanian, A.; Chandraker, M.; and Jawahar, C. V. 2022. IDD-3D: Indian Driving Dataset for 3D Unstructured Road Scenes. arXiv:2210.12878.

Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; and Koltun, V. 2017. CARLA: An Open Urban Driving Simulator. In Lopes, M.; Tamo, V.; and Abbeel, P., eds., *Proceedings of the 1st Annual Conference on Robot Learning*, volume 78 of *Proceedings of Machine Learning Research*, 1–16. PMLR.

Hu, E.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685*.

Hu, Y.; Yang, J.; Chen, L.; Li, K.; Sima, C.; Zhu, X.; Chai, S.; Du, S.; Lin, T.; Wang, W.; Lu, L.; Jia, X.; Liu, Q.; Dai, J.; Qiao, Y.; and Li, H. 2023. Planning-oriented Autonomous Driving. arXiv:2212.10156.

Janai, J.; Güney, F.; Behl, A.; and Geiger, A. 2020. Computer Vision for Autonomous Vehicles: Problems, Datasets and State-of-the-Art. *Foundations and Trends® in Computer Graphics and Vision*, 12(1-2): 1–308.

Kirkpatrick, J.; Pascanu, R.; Botvinick, M.; Blundell, C.; Veness, J.; Kim, G.; and Hassabis, D. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13): 3521–3526.

Liang, P.; Bommasani, R.; Lee, T.; Tsipras, D.; Mackey, H. S.; Zhang, C.; Tseng, E.; Hallinan, H.; Tham, L.; Koreeda, Y.; et al. 2022. Holistic Evaluation of Language Models. *arXiv preprint arXiv:2211.09110*.

Liu, Y.; Li, T.; Huang, Z.; Yang, Z.; and Huang, X. 2025. Bi-LoRA: Efficient Sharpness-Aware Minimization for Fine-Tuning Large-Scale Models. *arXiv preprint arXiv:2508.19564*.

McCloskey, M.; and Cohen, N. J. 1989. Catastrophic interference in connectionist networks. *Psychological review*, 96(4): 646.

Mukhoti, J.; Gal, Y.; Torr, P. H. S.; and Dokania, P. K. 2024. Fine-tuning can cripple your foundation model; preserving features may be the solution. arXiv:2308.13320.

OpenAI. 2023. GPT-4V(ision) System Card. https://openai.com/index/gpt-4v-system-card/.

Qiao, Z.; Li, H.; Cao, Z.; and Liu, H. X. 2025. LightEMMA: Lightweight End-to-End Multimodal Model for Autonomous Driving. *arXiv preprint arXiv:2505.00284*.

Sima, C.; Renz, K.; Chitta, K.; Chen, L.; Zhang, H.; Xie, C.; Beißwenger, J.; Luo, P.; Geiger, A.; and Li, H. 2025. DriveLM: Driving with Graph Visual Question Answering. arXiv:2312.14150.

Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; and Hauth, A. 2025a. Gemini: A Family of Highly Capable Multimodal Models. arXiv:2312.11805.

Team, G.; Kamath, A.; Ferret, J.; Pathak, S.; Vieillard, N.; Merhej, R.; Perrin, S.; Matejovicova, T.; and Ramé, A. 2025b. Gemma 3 Technical Report. arXiv:2503.19786.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.

Wang, L.; Zhang, X.; and Zhu, J. 2024. A Comprehensive Survey of Continual Learning: Theory, Method and Application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10): 7065–7088.

Wigness, M. B.; Eum, S.; Rogers, J. G.; Han, D.; and Kwon, H. 2019. A RUGD Dataset for Autonomous Navigation and Visual Perception in Unstructured Outdoor Environments. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 5000–5007.

Wilson, B.; Qi, W.; Agarwal, T.; Lambert, J.; Singh, J.; Khandelwal, S.; Pan, B.; Kumar, R.; Hartnett, A.; Pontes, J. K.; Ramanan, D.; Carr, P.; and Hays, J. 2023. Argoverse 2: Next Generation Datasets for Self-Driving Perception and Forecasting. arXiv:2301.00493.

Xu, Z.; Zhang, Y.; Xie, E.; Zhao, Z.; Guo, Y.; Wong, K.-Y. K.; Li, Z.; and Zhao, H. 2024. DriveGPT4: Interpretable End-to-end Autonomous Driving via Large Language Model. arXiv:2310.01412.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; and Gao, C. 2025. Qwen3 Technical Report. arXiv:2505.09388.

Yu, F.; Chen, H.; Wang, X.; Xian, W.; Chen, Y.; Liu, F.; Madhavan, V.; and Darrell, T. 2020. BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning. arXiv:1805.04687.

Zheng, W.; Song, R.; Guo, X.; Zhang, C.; and Chen, L. 2024. GenAD: Generative End-to-End Autonomous Driving. arXiv:2402.11502.

Zhou, X.; Liu, M.; Yurtsever, E.; Zagar, B. L.; Zimmer, W.; Cao, H.; and Knoll, A. C. 2024. Vision Language Models in Autonomous Driving: A Survey and Outlook. arXiv:2310.14414.

Reproducibility Checklist

1. General Paper Structure

1.1. Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes/partial/no/NA) yes

1.2. Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes/no) yes

1.3. Provides well-marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper (yes/no) yes

2. Theoretical Contributions

2.1. Does this paper make theoretical contributions? (yes/no) no

If yes, please address the following points:

2.2. All assumptions and restrictions are stated clearly and formally (yes/partial/no)

2.3. All novel claims are stated formally (e.g., in theorem statements) (yes/partial/no)

2.4. Proofs of all novel claims are included (yes/partial/no)

2.5. Proof sketches or intuitions are given for complex and/or novel results (yes/partial/no)

2.6. Appropriate citations to theoretical tools used are given (yes/partial/no)

2.7. All theoretical claims are demonstrated empirically to hold (yes/partial/no/NA)

2.8. All experimental code used to eliminate or disprove claims is included (yes/no/NA)

3. Dataset Usage

3.1. Does this paper rely on one or more datasets? (yes/no) yes

If yes, please address the following points:

3.2. A motivation is given for why the experiments are conducted on the selected datasets (yes/partial/no/NA) yes

3.3. All novel datasets introduced in this paper are included in a data appendix (yes/partial/no/NA) partial

3.4. All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no/NA) yes

3.5. All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations (yes/no/NA) yes

3.6. All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available (yes/partial/no/NA) yes

3.7. All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing (yes/partial/no/NA) NA

4. Computational Experiments

4.1. Does this paper include computational experiments? (yes/no) yes

If yes, please address the following points:

4.2. This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting (yes/partial/no/NA) yes

4.3. Any code required for pre-processing data is included in the appendix (yes/partial/no) no

4.4. All source code required for conducting and analyzing the experiments is included in a code appendix (yes/partial/no) no

4.5. All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no) partial

4.6. All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes/partial/no) no

4.7. If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results (yes/partial/no/NA) yes

4.8. This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks (yes/partial/no) yes

4.9. This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics (yes/partial/no) partial

4.10. This paper states the number of algorithm runs used to compute each reported result (yes/no) yes

4.11. Analysis of experiments goes beyond singledimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information (yes/no) no

4.12. The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank) (yes/partial/no) no

4.13. This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments (yes/partial/no/NA) yes