

SF-V: Auditing Soft Headway-Floor Regularization at the VLM-to-MPC Interface

Giwoong Kim^{1,2}, Junbeom Park^{1,2}, and Shiho Kim¹

¹ School of Integrated Technology, Yonsei University, Seoul, Republic of Korea

² BK21 Graduate Program in Intelligent Semiconductor Technology, Yonsei University, Seoul, Republic of Korea
{7190,bumy0627,shiho}@yonsei.ac.kr

Abstract. VLM-generated parameters can directly alter downstream MPC behavior. We audit desired headway below a controller-side 2 s reference as a configuration-specific interface event, not a universal safety violation. We refer to this fixed-target soft-floor intervention as SF-V (Safety-Floor VLM-MPC). SF-V adds a fixed soft headway-floor term to a reduced longitudinal MPC scaffold; eliminating its unbounded slack gives an exact squared-hinge penalty with no feasible-set change or hard guarantee. Of 303 logged outputs, 66 are sub-floor and headway has weak rank association with logged gap and closing-time proxies. Controlled cost-only, clamp, matched-target, and SF-V comparisons separate clipping, penalty strength, and target choice. SF-V reduces mean post-initial worst deficit by 0.126/0.617/0.470 m in clean/lead/TTC settings. Exact average-speed conditioning retains 0.033/0.166/0.150 m, but independent margin gains are small and extreme-tail and collision outcomes do not improve. A 4 s braking diagnostic increases mean minimum gap by 0.668 m without lag and 0.579 m with 0.5 s lag, again without collision-count change. This is a bounded interface audit, not deployment-level safety validation.

Keywords: Vision-language model · Model predictive control · Autonomous driving · Headway regularization · Defensive driving

1 Introduction

Vision-language models (VLMs) connect scene understanding to planning and control [12–14, 22]. VLM-MPC generates the prediction horizon, cost weights, desired speed, and desired headway for downstream MPC [6, 11]; aggressive outputs can therefore create interface risk, particularly in defensive-driving tails [4, 18, 19].

We study $h_d < h_{\text{floor}}$ as a controller-relative sub-floor event, not a universal declaration that $h_d < 2$ s is unsafe. The contribution is a three-level audit: (i) parameter-level frequency and risk-proxy association; (ii) controller-level isolation of clipping, penalty, target, and average-speed effects under identical logged outputs; and (iii) outcome-level gap/TTC, tail, and collision analysis. SF-V is the fixed-target probe, not a new safety architecture. We derive its squared-hinge form and test sensitivity and sudden braking.

2 Related Work

VLM driving systems cover structured reasoning, language-conditioned control, and hierarchical planning [12–14]. LanguageMPC, ChatMPC, and VLM-UDMC map language or vision-language output to controller decisions [5, 11, 20]; VLM-MPC is closest because it outputs six MPC parameters [6]. VisioPath and concurrent VLM-CASE add optimization-side verification or adaptive envelopes [17, 21]. Our narrower scope is the logged VLM-to-MPC headway interface.

Longitudinal MPC commonly uses constant-time-headway spacing [7, 16]. Hard constraints restrict admissible trajectories; softened constraints may retain feasibility [10]. CBF-QPs and predictive safety filters instead modify nominal actions against explicit conditions [1, 2, 15]. SF-V is only an in-objective soft penalty: it provides neither forward invariance nor collision avoidance and is not compared with verified filters.

3 Methodology

SF-V retains the six-parameter interface but uses a reduced gap–speed scaffold, not the full published VLM-MPC dynamics [6], to isolate one downstream objective term (Fig. 1).

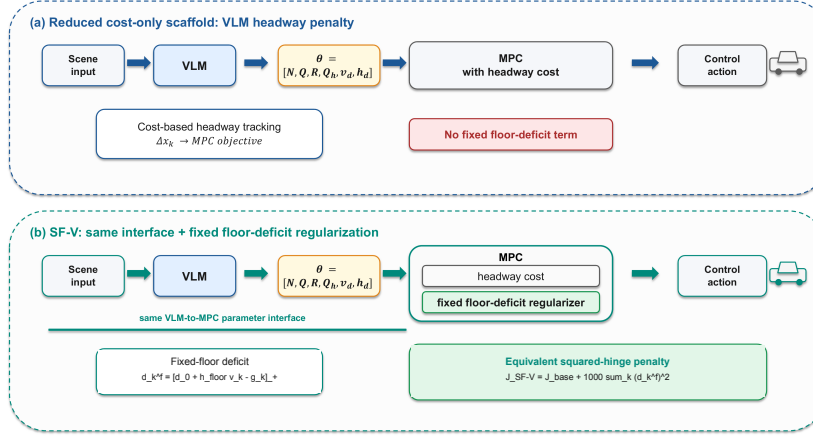


Fig. 1: Controlled comparison with one shared logged VLM vector. SF-V adds only a fixed squared-hinge floor-deficit term; stress tests alter the initial gap without requerying the VLM.

The state $\mathbf{s}_k = [g_k, v_k]^\top$ contains lead gap and ego speed; u_k is acceleration and lead speed is constant within each rollout:

$$g_{k+1} = g_k + \Delta t(v_{\text{lead}} - v_k), \quad (1)$$

$$v_{k+1} = v_k + \Delta t u_k. \quad (2)$$

The VLM vector is $\theta = [N, Q, R, Q_h, v_d, h_d]$ [6]. With $[z]_+ = \max(0, z)$ and $e_k^g = [h_d v_k + d_0 - g_k]_+$, the reduced cost-only objective is

$$J_{\text{base}} = \sum_{k=0}^{\bar{N}-1} [Q(v_k - v_d)^2 + R u_k^2 + Q_h (e_k^g)^2] + 2Q_h (e_N^g)^2. \quad (3)$$

SF-V sets $d_0 = 5.0$ m, $h_{\text{floor}} = 2.0$ s, and $d_k^f = [d_0 + h_{\text{floor}} v_k - g_k]_+$. Its implementation imposes $g_k + \epsilon_k \geq d_0 + h_{\text{floor}} v_k$, $\epsilon_k \geq 0$, and penalizes $1000 \sum_k \epsilon_k^2$. Because slack is unbounded, $\epsilon_k^* = d_k^f$, yielding exactly

$$J_{\text{SF-V}} = J_{\text{base}} + 1000 \sum_{k=0}^{\bar{N}} (d_k^f)^2. \quad (4)$$

Thus the feasible set is unchanged: SF-V is a squared-hinge regularizer, not a hard constraint or safety guarantee. The $k = 0$ term is decision-constant, and all fixed values are engineering choices rather than universal thresholds.

Mechanism Baselines. Clamp- Q_h replaces h_d in Eq. (3) by $\max(h_d, 2.0)$ s while retaining native Q_h . To separate target choice from added weight, VLM-target uses

$$J_{\text{VLM-target}} = J_{\text{base}} + 1000 \sum_{k=0}^{\bar{N}} [d_0 + h_d v_k - g_k]_+^2. \quad (5)$$

VLM-target versus SF-V isolates fixed target choice; Clamp- Q_h additionally separates high weight from clipping. VLM-target and SF-V coincide when $h_d = 2.0$ s.

4 Experimental Setup

We use paired one-shot rollouts with constant lead speed and a separate receding-horizon braking diagnostic; neither is interactive simulation. Stored parameters come from `llava-hf/llava-v1.6-mistral-7b-hf` applied once to each front image and derived fields. A fixed prompt requests two scene/risk cues and six numbers; greedy decoding uses at most 220 tokens. All methods receive identical parsed vectors, and stress tests do not re-query the VLM. Results are therefore conditional on this single checkpoint, prompt, and decoding path. CasADi 3.7.2 with IPOPT solves every reported problem.

Logged Output Audit. For 303 clean logs, $h_d \in \{1.0, 1.5, 2.0, 3.0\}$ s has counts 16, 50, 230, 7; 66 outputs are sub-floor. Desired speed spans 0.01–13.57 m/s (median 10.0); $Q_h \in \{0.1, 0.5, 1.0, 5.0\}$ has counts 3, 21, 265, 14. Hence $\lambda/Q_h = 200$ –10,000 (median 1000), a coefficient comparison rather than total objective influence because targets and terminal scaling differ. Every output requests $N = 10$, capped at $\bar{N} = 8$. Among 145 detected leads, 34 are sub-floor; 9/40 logs with

positive-closing $\text{TTC}_{\log} = g_0/(v_0 - v_{\text{lead}}) \leq 8 \text{ s}$ use $h_d = 1.5 \text{ s}$. Spearman associations of h_d with speed, gap, and positive-closing $\text{TTC}_{\log}$ are -0.11 , 0.02 , and -0.11 , respectively: weak configuration-specific evidence, not proof of VLM error.

Configuration. We use $\Delta t = 0.2 \text{ s}$, $\bar{N} = \max(1, \min(N, 8))$, $0 \leq v_k \leq 25 \text{ m/s}$, and $-4 \leq u_k \leq 2 \text{ m/s}^2$. All methods otherwise share inputs and bounds. One-factor sensitivity varies floor $1.5/2.0/2.5 \text{ s}$, weight $10^2/10^3/10^4$, and cap $4/8/10$, anchored at $(2.0, 10^3, 8)$.

Speed Conditioning. We re-solve cost-only subject to

$$\frac{1}{\bar{N} + 1} \sum_{k=0}^{\bar{N}} v_k = \bar{v}_{\text{SF-V}}, \quad (6)$$

with $\bar{v}_{\text{SF-V}}$ the realized SF-V mean speed. All 303/126/63 solves succeed with maximum error below $9 \times 10^{-11} \text{ m/s}$. Because this condition uses the evaluated outcome, it separates average slowdown only as a post-hoc diagnostic, not a deployable baseline or causal proof.

Scenarios. Clean uses all 303 logged states [3]; no-lead scenes receive a disclosed inactive 60 m surrogate. Counterfactual stress sets $g_0^{\text{stress}} = \max(5 \text{ m}, 0.5g_0^{\log})$. Lead stress contains 126 scenes with gap $\leq 20 \text{ m}$ or defined $\text{TTC} \leq 8 \text{ s}$; TTC stress contains 63 positive-closing scenes with $g_0^{\text{stress}}/(v_0 - v_{\text{lead}}) \leq 8 \text{ s}$ [8]. Sub-floor counts are 66/32/18 in clean/lead/TTC; all other stress outputs use $h_d = 2 \text{ s}$.

Sudden-Braking Diagnostic. The 63 TTC states are re-solved every 0.2 s for 4 s while the lead brakes at -4 m/s^2 after 0.4 s. Each solve assumes current lead speed is constant; a second profile adds ego-actuation lag $\tau = 0.5 \text{ s}$. We report minimum gap, positive-closing relative TTC, collision, and a disclosed 2 m near-collision count. This remains a reduced-model diagnostic.

Metrics. For $\mathcal{K}_+ = \{k : g_k > 0, v_k > 0.1 \text{ m/s}\}$, we use

$$T_{\text{GoS}} = \min_{k \in \mathcal{K}_+} g_k / v_k, \quad D_{\text{mean}} = \frac{1}{\bar{N}} \sum_{k=1}^{\bar{N}} d_k^{\text{f}}, \quad (7)$$

$$D_{\text{worst}} = \max_{1 \leq k \leq \bar{N}} d_k^{\text{f}}, \quad \text{RMSa} = \left[\frac{1}{\bar{N}} \sum_{k=0}^{\bar{N}-1} \left(\frac{v_{k+1} - v_k}{\Delta t} \right)^2 \right]^{1/2}. \quad (8)$$

GoS is zero on moving collision and undefined if $\mathcal{K}_+$ is empty; it is a same-lane proxy, not geometric PET [9]. Deficits exclude uncontrollable $k = 0$, and RMSa is acceleration effort, not jerk. GoS uses 296/118/63 common valid scenes; other metrics use 303/126/63. Paired intervals use 10,000 bootstrap resamples.

5 Evaluation

Table 1: One-shot means. GoS uses common $n = 296/118/63$; deficits exclude $k = 0$.

Setting	Method	n	Speed	MinDist $\uparrow$	GoS $\uparrow$	RMSa $\downarrow$	MeanDef $\downarrow$	WorstDef $\downarrow$
Clean	Cost-only	303	5.70	37.72	8.16	1.19	0.33	0.56
Clean	Clamp- Q_h	303	5.69	37.72	8.17	1.20	0.30	0.54
Clean	VLM-target	303	5.67	37.76	8.26	1.16	0.25	0.44
Clean	SF-V	303	5.66	37.76	8.27	1.17	0.22	0.43
Lead stress	Cost-only	126	3.73	9.40	3.55	1.89	1.97	3.12
Lead stress	Clamp- Q_h	126	3.68	9.42	3.59	1.94	1.82	2.99
Lead stress	VLM-target	126	3.59	9.57	4.60	1.81	1.56	2.63
Lead stress	SF-V	126	3.53	9.59	4.65	1.86	1.42	2.50
TTC stress	Cost-only	63	4.75	9.99	2.43	2.24	2.51	3.64
TTC stress	Clamp- Q_h	63	4.72	10.03	2.45	2.28	2.45	3.49
TTC stress	VLM-target	63	4.65	10.15	2.46	2.28	2.24	3.33
TTC stress	SF-V	63	4.63	10.18	2.49	2.31	2.19	3.17

Mechanism Isolation. SF-V reduces WorstDef over cost-only by 0.126, 0.617, and 0.470 m in clean, lead, and TTC settings. Clamp- Q_h yields 0.011, 0.125, and 0.153 m; VLM-target yields 0.115, 0.485, and 0.310 m. Fixing the high-weight target adds 0.011, 0.132, and 0.161 m, with intervals $[0.000, 0.031]$, $[0.049, 0.233]$, and $[0.033, 0.320]$, and never worsens WorstDef. High weight explains most aggregate change; clipping alone is insufficient.

Speed Control. At matched speed, SF-V retains positive WorstDef reductions. The clean reduction is 0.033 m (95% interval $[0.020, 0.047]$); lead and TTC reductions are 0.166 m ($[0.120, 0.215]$) and 0.150 m ($[0.100, 0.203]$). Minimum-gap gains are only 0.003, 0.012, and 0.016 m; GoS gains are 0.013, 0.043, and 0.001 s, with stress intervals crossing zero. SF-V reshapes its aligned deficit beyond average slowdown, but independent margin evidence is small.

Table 2: Subgroup WorstDef reduction [95% interval], m. Lead/floor compare cost-only to SF-V; $h_d < 2$ s compares VLM-target to SF-V.

Subgroup	Setting	n	Reduction [95% interval]
Lead-present	Clean	145	0.264 [0.174, 0.363]
	Lead stress	126	0.617 [0.476, 0.764]
	TTC stress	63	0.470 [0.301, 0.660]
Floor-active	Clean	52	0.736 [0.534, 0.948]
	Lead stress	108	0.720 [0.568, 0.881]
	TTC stress	57	0.520 [0.341, 0.725]
$h_d < 2$ s	Clean	66	0.051 [0.000, 0.146]
	Lead stress	32	0.520 [0.220, 0.873]
	TTC stress	18	0.562 [0.164, 1.056]

Floor-active denotes cost-only WorstDef $> 10^{-6}$ m; lead/TTC sets are lead-present. The clean fixed-target interval includes zero, while lead/TTC intervals do not.

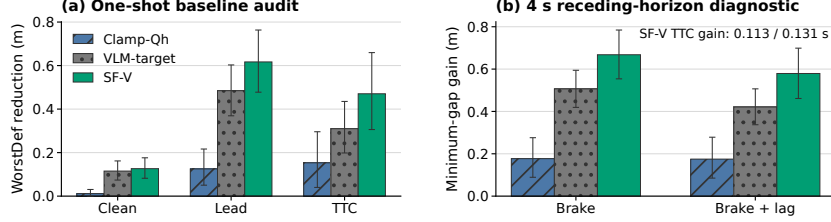


Fig. 2: Validation versus cost-only: (a) WorstDef reduction; (b) 4 s braking minimum-gap gain. Error bars are paired-bootstrap 95% intervals.

Weights 10^2 – 10^4 change reduction by at most 0.007 m. Raising the floor from 1.5 to 2.0 s changes clean/lead/TTC reductions from 0.063/0.299/0.083 to 0.126/0.617/0.470 m; 2.5 s saturates deficit reduction while slowing further. Caps 4/8/10 retain positive intervals on a common 0.8 s window.

In Fig. 2(b), SF-V raises braking minimum gap by 0.668 m without lag (interval $[0.554, 0.784]$) and 0.579 m with lag ($[0.461, 0.699]$); relative-TTC gains are 0.113 and 0.131 s.

Tail and Paired Collision Audit. Cost-only-to-SF-V 95th-percentile WorstDef changes $3.625 \rightarrow 3.599$ m in clean data but remains 11.897/13.272 m under lead/TTC stress; cost-only-defined upper-5% paired reductions are 0.006/0/0 m. One-shot collisions remain 1/4/4 with every paired identity unchanged; speed conditioning gives the same p95 and identities. Braking collisions are 4 without lag and 5 with lag for both methods, with 6/11 near collisions. Mean margins improve, but extreme stress tails and observed collisions do not.

Runtime. After six warm-ups on one Xeon Gold 6230 process, cost-only mean, median, p95, and maximum runtime is 27.25, 27.05, 29.58, and 35.72 ms; SF-V takes 46.70, 46.91, 50.55, and 68.81 ms. All 303 solves finish within 200 ms; this is hardware-specific, not a hard real-time guarantee.

Limitations. This reduced, logged-output study uses one checkpoint/prompt, no VLM re-query, a surrogate no-lead gap, and scripted rather than interactive traffic. Optimized deficits improve after speed conditioning, but independent margins are small and tail/collision outcomes do not change; the conditioner itself is post hoc. One-factor sensitivities and server runtime provide no robustness or real-time certificate. Without hard-MPC or verified-filter baselines, no formal, deployment-level, or superiority claim follows.

6 Conclusion

SF-V shows that a high-weight fixed target can reduce mean floor deficit beyond clipping and average slowdown. Yet independent gains are small and collision identity and extreme tails do not improve. The result is a bounded audit of soft VLM-to-MPC intervention, not a safety guarantee; future work requires VLM re-querying, cross-model tests, interactive validation, and verified-filter comparisons.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2025-02218237, Development of Digital Innovative Technologies for Enhancing the Safety of Complex Autonomous Mobility).

References

1. Ames, A.D., Grizzle, J.W., Tabuada, P.: Control barrier function based quadratic programs with application to adaptive cruise control. In: 53rd IEEE Conference on Decision and Control. pp. 6271–6278 (2014). <https://doi.org/10.1109/CDC.2014.7040372>
2. Ames, A.D., Xu, X., Grizzle, J.W., Tabuada, P.: Control barrier function based quadratic programs for safety critical systems. *IEEE Transactions on Automatic Control* **62**(8), 3861–3876 (2017). <https://doi.org/10.1109/TAC.2016.2638961>
3. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuScenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11621–11631 (2020). <https://doi.org/10.1109/CVPR42600.2020.01164>
4. Hanselmann, N., Renz, K., Chitta, K., Bhattacharyya, A., Geiger, A.: KING: Generating safety-critical driving scenarios for robust imitation via kinematics gradients. In: Computer Vision – ECCV 2022. Lecture Notes in Computer Science, vol. 13698, pp. 335–352. Springer (2022). https://doi.org/10.1007/978-3-031-19839-7_20
5. Liu, H., Guo, H., Liu, P., Ma, B., Zhang, Y., Ma, J., Lee, T.H.: VLM-UDMC: Vlm-enhanced unified decision-making and motion control for urban autonomous driving. *arXiv preprint arXiv:2507.15266* (2025). <https://doi.org/10.48550/arXiv.2507.15266>
6. Long, K., Shi, H., Liu, J., Xiao, C., Li, X.: VLM-MPC: Model predictive controller augmented vision language model for autonomous driving. *Transportation Research Part C: Emerging Technologies* **183**, 105487 (2026). <https://doi.org/10.1016/j.trc.2025.105487>
7. Mayne, D.Q., Rawlings, J.B., Rao, C.V., Sokaert, P.O.M.: Constrained model predictive control: Stability and optimality. *Automatica* **36**(6), 789–814 (2000). [https://doi.org/10.1016/S0005-1098\(99\)00214-9](https://doi.org/10.1016/S0005-1098(99)00214-9)

8. Minderhoud, M.M., Bovy, P.H.L.: Extended time-to-collision measures for road traffic safety assessment. *Accident Analysis & Prevention* **33**(1), 89–97 (2001). [https://doi.org/10.1016/S0001-4575\(00\)00019-1](https://doi.org/10.1016/S0001-4575(00)00019-1)
9. Peesapati, L.N., Hunter, M.P., Rodgers, M.O.: Can post encroachment time substitute intersection characteristics in crash prediction models? *Journal of Safety Research* **66**, 205–211 (2018). <https://doi.org/10.1016/j.jsr.2018.05.002>
10. Scokaert, P.O.M., Rawlings, J.B.: Feasibility issues in linear model predictive control. *AIChE Journal* **45**(8), 1649–1659 (1999). <https://doi.org/10.1002/aic.690450805>
11. Sha, H., Mu, Y., Jiang, Y., Chen, L., Xu, C., Luo, P., Li, S.E., Tomizuka, M., Zhan, W., Ding, M.: LanguageMPC: Large language models as decision makers for autonomous driving. *arXiv preprint arXiv:2310.03026* (2023). <https://doi.org/10.48550/arXiv.2310.03026>
12. Shao, H., Hu, Y., Wang, L., Song, G., Waslander, S.L., Liu, Y., Li, H.: LMDrive: Closed-loop end-to-end driving with large language models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 15120–15130 (2024). <https://doi.org/10.1109/CVPR52733.2024.01432>
13. Sima, C., Renz, K., Chitta, K., Chen, L., Zhang, H., Xie, C., Beißwenger, J., Luo, P., Geiger, A., Li, H.: DriveLM: Driving with graph visual question answering. In: *Computer Vision – ECCV 2024*. pp. 256–274. Springer (2024). https://doi.org/10.1007/978-3-031-72943-0_15
14. Tian, X., Gu, J., Li, B., Liu, Y., Wang, Y., Zhao, Z., Zhan, K., Jia, P., Lang, X., Zhao, H.: DriveVLM: The convergence of autonomous driving and large vision-language models. In: *Proceedings of the 8th Conference on Robot Learning. Proceedings of Machine Learning Research*, vol. 270, pp. 4698–4726. PMLR (2025), <https://proceedings.mlr.press/v270/tian25c.html>
15. Wabersich, K.P., Zeilinger, M.N.: A predictive safety filter for learning-based control of constrained nonlinear dynamical systems. *Automatica* **129**, 109597 (2021). <https://doi.org/10.1016/j.automatica.2021.109597>
16. Wang, J., Rajamani, R.: Should adaptive cruise-control systems be designed to maintain a constant time gap between vehicles? *IEEE Transactions on Vehicular Technology* **53**(5), 1480–1490 (2004). <https://doi.org/10.1109/TVT.2004.832386>
17. Wang, S., Typaldos, P., Li, C., Malikopoulos, A.A.: VisioPath: Vision-language enhanced model predictive control for safe autonomous navigation in mixed traffic. *IEEE Open Journal of Control Systems* **4**, 562–580 (2025). <https://doi.org/10.1109/OJCSYS.2025.3620149>
18. Xie, S., Kong, L., Dong, Y., Sima, C., Zhang, W., Chen, Q.A., Liu, Z., Pan, L.: Are VLMs ready for autonomous driving? an empirical study from the reliability, data, and metric perspectives. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 6585–6597 (2025). <https://doi.org/10.1109/ICCV51701.2025.00621>
19. Xu, C., Ding, W., Lyu, W., Liu, Z., Wang, S., He, Y., Hu, H., Zhao, D., Li, B.: SafeBench: A benchmarking platform for safety evaluation of autonomous vehicles. In: *Advances in Neural Information Processing Systems*. vol. 35, pp. 25667–25682 (2022). <https://doi.org/10.52202/068431-1861>
20. Xu, W., Yin, Z., Zhou, Y., Zhu, Y., Wang, M., Lei, J., Chen, H.: ChatMPC: A language-driven model predictive control framework for adaptive and personalized autonomous driving. *Autonomous Intelligent Systems* **5**(1), 27 (2025). <https://doi.org/10.1007/s43684-025-00116-x>

21. Yang, T., Li, K., Qin, R., Hu, X.: VLM-CASE: Vision-language model enabled context-adaptive safety envelopes for anticipatory safe autonomous driving. arXiv preprint arXiv:2607.05180 (2026). <https://doi.org/10.48550/arXiv.2607.05180>
22. Zhou, X., Liu, M., Yurtsever, E., Zagar, B.L., Zimmer, W., Cao, H., Knoll, A.C.: Vision language models in autonomous driving: A survey and outlook. IEEE Transactions on Intelligent Vehicles pp. 1–20 (2024). <https://doi.org/10.1109/TIV.2024.3402136>