RiskLens: Auditing the Causal Geometry of End-to-End Driving Policies

Ziyuan Yang, Zengxiang Lei, Satish V. Ukkusuri*
Lyles School of Civil and Construction Engineering
Purdue University
West Lafayette, IN, United States
{yang2995, lei67, sukkusur}@purdue.edu

***Abstract*—End-to-end driving policies have achieved strong empirical performance, yet it remains unclear whether their decisions reflect causal driving principles or only correlate scene patterns with actions. A simplified form of this question is whether a policy can understand the risk level of the current scenario and adjust its behavior accordingly, such as becoming more conservative under higher risk. We introduce RiskLens, a causal auditing framework for examining the internal causal geometry of end-to-end driving policies. RiskLens audits three representative policy families, including modular, vision-language, and diffusion-based driving models, by testing whether risk is decodable from hidden states, organized into intervention-effective subspaces, and propagated to downstream behavior. We find pervasive "risk amnesia" in end-to-end driving models—most pronounced in VLM architectures—where linearly decodable risk representations become causally decoupled from final driving decisions in later layers.**

***Index Terms*—End-to-end autonomous driving, safety auditing, risk representation, model evaluation.**

I. Introduction

End-to-end driving policies have recently emerged as a leading paradigm in autonomous driving, achieving strong performance across planning and control benchmarks [1]–[3]. By mapping high-dimensional scene observations directly to trajectories or actions, these models offer strong scalability and empirical effectiveness. However, strong behavioral performance alone does not show whether a policy has learned causal driving principles, or merely correlates scene patterns with actions.Prior work on causal confusion suggests that policies may exploit spurious correlations or statistical shortcuts rather than genuine safety mechanisms [4], [5], raising particular concern in rare and safety-critical scenarios. This motivates our central question: **Can end-to-end driving policies understand the risk level of a scenario and adjust their behavior accordingly?**

Recent work has introduced causality into autonomous driving through de-confounding methods [6]–[8], causal inductive biases and planning objectives [9]–[12], and vision-language systems that connect causal reasoning with action spaces [13], [14]. These methods mainly use causality as a forward-design principle for improving robustness and generalization. In parallel, causal frameworks have been used for safety auditing and high-risk scenario generation [15], [16], but they typically treat the driving policy as a black box. As a result, they do not directly reveal whether an already trained policy internally represents traffic-relevant risk, organizes it into an intervention-effective structure, or propagates it to downstream behavior.

In this work, we introduce **RiskLens**, a model-agnostic causal auditing framework for examining the internal risk-to-behavior chain of end-to-end driving policies. Rather than treating policy behavior as a black box, RiskLens analyzes how risk-related information is represented, organized into intervention-relevant latent directions, and ultimately translated into driving actions. Across modular, vision-language, and diffusion-based driving policies, RiskLens localizes different failure modes: missing risk information, non-causal latent organization, or a broken interface between risk representation and action generation.

Our contributions are threefold. First, we formulate causal risk auditing as a diagnostic problem for end-to-end driving policies, shifting safety evaluation from output-level performance to the internal risk-to-behavior chain. Second, we adapt probing, low-rank subspace discovery, and causal intervention into a staged audit of internal risk representations. Third, we provide comparative evidence showing that risk can be clearly decodable from hidden states while still failing to induce conservative behavior, revealing a form of risk amnesia in end-to-end driving policies.

II. Related Work

A. Causality in Autonomous Driving

End-to-end driving models are typically trained on large-scale driving logs, making them vulnerable to spurious correlations between scene context and driving behavior. Causal confusion, where policies exploit statistical shortcuts rather than true causal mechanisms, is a well-documented failure mode in imitation learning [4], especially under the closed-loop dynamics of autonomous driving [5]. Recent end-to-end and planning-oriented systems have improved driving performance through conditional imitation, sensor fusion, vectorized planning, modular planning heads, and diffusion-based planning [1]–[3], [17]–[19]. However, strong benchmark

*Corresponding author: Satish V. Ukkusuri (sukkusur@purdue.edu).

performance does not reveal whether a model internally uses safety-relevant risk variables.

Recent work has introduced causality into trajectory prediction, planning, and decision-making. Some methods mitigate confounding during training, including CausalVAD [6], Prioritizing Perception-Guided Self-Supervision [7], Deconfounded Lifelong Learning [8], and the Causal Awareness Framework [11]. Others incorporate causal structure into representation learning, planning, and generalization, such as CaDeT [20], CauTraj [10], PCM [12], CRiTIC [9], Echo Planning [21], invariant learning [22], and causal representation learning [23]. Recent vision-language driving systems further connect causal reasoning with action spaces through instructed action generation and causal scene narration [13], [14].

Despite these advances, most studies use causality as a forward-design principle, such as an inductive bias, architectural constraint, training signal, or planning objective. Related safety-auditing methods, including DNN testing, scenario generation, simulation platforms, and safety benchmarks [15], [16], [24]–[29], often treat the policy as a black box and focus on system-level failures. These approaches expose unsafe behaviors but do not examine whether an already trained end-to-end model has internalized traffic-relevant causal structure. In contrast, RiskLens audits whether learned policies encode, organize, and execute risk-related information after training.

B. Causal Analysis in Deep Learning Models

Our work also connects to causal analysis, concept-based interpretability, and mechanistic interpretability in deep learning. Classic causal inference and causal representation learning provide the basis for linking high-level variables to learned representations [23], [30], [31]. In neural networks, attribution and concept-based methods such as LIME, Grad-CAM, Network Dissection, and TCAV explain predictions or identify human-interpretable concepts [32]–[35]. More recent mechanistic methods study internal computation through causal tracing, mediation analysis, circuits, and representation interventions [36]–[40].

Related intervention-based analyses have been extended to reinforcement learning [41], [42], multi-agent systems [43], and generative vision models, including diffusion-model bias analysis, concept evolution across denoising steps, and interpretable feature extraction in diffusion transformers [44]–[46]. Applying these tools to autonomous driving requires domain grounding, because driving risk is continuous, interactive, and physically expressed through conflict geometry, clearance, time-to-conflict, trajectory adjustment, agent intent, and scenario context [47]–[50]. We address this gap by grounding causal auditing in safety-relevant risk variables and testing not only whether risk is decodable, but also whether it forms intervention-effective subspaces and propagates to behavior.

III. Problem Setup

A. End-to-End Driving Policies and Internal Representations

We study end-to-end driving policies that map an input scenario $x \in \mathcal{X}$ to a future ego plan. The scenario may include sensor observations, maps, ego states, surrounding agents, language tokens, or other model-specific inputs. A policy π_θ predicts a future trajectory or action sequence, and its hidden representation at an internal site ℓ is denoted as

$$\tau = \pi_\theta(x), \quad \tau \in \mathbb{R}^{T\times d_a}, \qquad \mathbf{h}_\ell(x) \in \mathbb{R}^{d_\ell}. \tag{1}$$

Here, ℓ may correspond to a decoder layer, transformer block, token position, fusion module, or denoising stage. This abstraction provides a common interface for auditing heterogeneous driving policies. We ask whether $\mathbf{h}_\ell(x)$ merely contains risk-correlated information, or whether it organizes such information in a causally meaningful geometry that can influence the final plan τ.

B. Transportation-Grounded Risk and Matched Scenario Pairs

RiskLens defines risk independently of the model being audited. For each scenario x, we assign a transportation-grounded risk score $r(x)$ using conflict-relevant quantities such as distance-to-conflict, time-to-conflict, relative motion, and agent proximity, and define a binary auditing label

$$r(x) \in \mathbb{R}_{\geq 0}, \qquad y(x) = \mathbb{I}[r(x) > \delta]. \tag{2}$$

To reduce confounding from road topology, maneuver type, traffic density, or ego kinematics, we construct matched safe–risky pairs with similar non-risk context $c(x)$ but different risk levels:

$$\begin{aligned} &\mathcal{P} = \{(x_i^{\text{safe}}, x_i^{\text{risk}})\}_{i=1}^N, \\ \text{s.t.} \quad &d_c\big(c(x_i^{\text{safe}}), c(x_i^{\text{risk}})\big) \leq \epsilon_c, \\ &r(x_i^{\text{risk}}) - r(x_i^{\text{safe}}) \geq \epsilon_r. \end{aligned} \tag{3}$$

Each pair induces a representation-level contrast

$$\Delta\mathbf{h}_\ell^{(i)} = \mathbf{h}_\ell(x_i^{\text{risk}}) - \mathbf{h}_\ell(x_i^{\text{safe}}), \tag{4}$$

which approximates a counterfactual risk contrast and supports intervention-based tests of whether high-level risk aligns with neural representations [39], [51]. Matching quality affects the strength of this contrast, so our conclusions are within-policy diagnostics rather than absolute cross-model rankings.

C. Causal Geometry Auditing Objective

RiskLens examines whether scenario-level risk is reflected in the internal computation of an end-to-end policy. Rather than modeling the full traffic causal graph, we focus on a minimal safety-relevant relation

$$R \longrightarrow B, \qquad R \longrightarrow \mathbf{h}_\ell \longrightarrow \tau, \tag{5}$$

where R denotes scenario-level risk and B denotes the policy's behavioral tendency. We refer to the organization of risk-related information in representation space as the policy's *causal geometry of risk*.

We audit this geometry through three criteria:

$$\begin{aligned} \text{Risk Decodability:} \quad & \hat{y}_\ell = q_\ell(\mathbf{h}_\ell(x)), \\ \text{Risk Geometry:} \quad & \mathcal{S}_\ell = \operatorname{span}(\mathbf{U}_\ell), \quad \mathbf{U}_\ell \in \mathbb{R}^{d_\ell \times k}, \\ \text{Behavioral Execution:} \quad & \Delta_\tau = D\Big(\pi_\theta(x; \tilde{\mathbf{h}}_\ell), \pi_\theta(x; \mathbf{h}_\ell)\Big). \end{aligned} \tag{6}$$

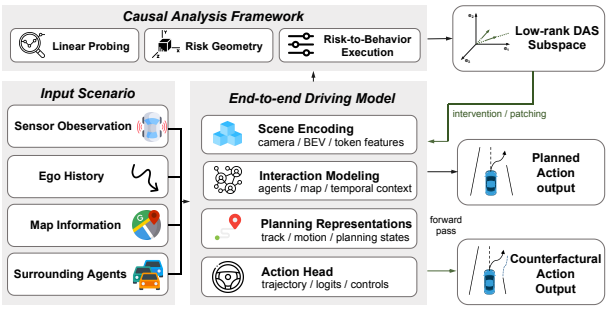

Fig. 1. **RiskLens risk-to-behavior auditing framework.** The pipeline separates risk availability, low-rank causal geometry, and downstream behavioral execution to localize where the internal risk-to-behavior chain fails.

These criteria distinguish three possible failure points: risk may be unavailable in the representation, available but not organized into an intervention-effective subspace, or steerable in latent space but filtered before final action.

IV. Methodology

A. Overview of Auditing Pipeline

RiskLens audits the internal risk-to-behavior chain through staged failure localization (Fig. 1). Linear probes are only an entry test for information availability; low-rank Distributed Alignment Search (DAS) [51] tests whether risk supports causal intervention; and held-out execution tests whether the intervened signal reaches logits or physical trajectories. Thus, later stages, not probe accuracy alone, determine causal use.

For the risk-availability stage, we train lightweight linear probes on frozen representations $\mathbf{h}_\ell(x)$ to predict the binary risk label $y(x)$ or correlate with the continuous risk score $r(x)$. We report AUROC for binary risk decodability and Spearman correlation for continuous risk association, using shuffled-label or random-feature controls where applicable. This stage is only an entry test for information availability; causal use is evaluated by the intervention and execution stages below.

B. Low-Rank Risk Geometry Discovery

For each matched pair, let $\mathbf{h}_b$ and $\mathbf{h}_s$ denote the base and source representations at site ℓ. DAS learns a rank-k orthonormal subspace

$$\mathcal{S}_\ell = \mathrm{span}(\mathbf{U}_\ell), \qquad \mathbf{U}_\ell \in \mathbb{R}^{d_\ell \times k}, \qquad \mathbf{U}_\ell^\top \mathbf{U}_\ell = \mathbf{I}. \tag{7}$$

Instead of replacing the full representation, RiskLens patches only the source–base difference projected onto this subspace:

$$\tilde{\mathbf{h}}_\ell = \mathbf{h}_b + \mathbf{U}_\ell \mathbf{U}_\ell^\top \left(\mathbf{h}_s - \mathbf{h}_b\right). \tag{8}$$

The subspace is optimized so that the downstream output after patching moves toward the source condition:

$$\min_{\mathbf{U}_\ell} \sum_{i=1}^{N} \mathcal{L}_{\mathrm{DAS}} \left(g_{\ell:\mathrm{out}}(\tilde{\mathbf{h}}_\ell^{(i)}), g_{\ell:\mathrm{out}}(\mathbf{h}_s^{(i)}) \right), \tag{9}$$

where $g_{\ell:\mathrm{out}}$ denotes the remaining computation from site ℓ to the output. We treat k as a validation-selected diagnostic hyperparameter within each policy and site; because dimensions and output interfaces differ, ranks are not used as direct cross-model capacity scores.

C. Risk-to-Behavior Execution

After learning $\mathbf{U}_\ell$, we evaluate whether the discovered risk geometry affects downstream behavior on held-out matched pairs. We compare the original output and the counterfactual output produced by the patched representation:

$$\tau = \pi_\theta(x; \mathbf{h}_\ell), \qquad \tilde{\tau} = \pi_\theta(x; \tilde{\mathbf{h}}_\ell), \qquad \Delta_\tau = D_\tau(\tilde{\tau}, \tau). \tag{10}$$

To quantify whether the intervention induces conservative driving behavior, we compute a Conservative Action Index (CAI). Let τ^L, τ^H, and $\tilde{\tau}$ denote the low-risk, high-risk, and patched trajectories. For each trajectory feature f_j with conservative orientation s_j (lower speed/progress: $s_j = -1$; stronger deceleration: $s_j = 1$), we compute

$$z_j = \frac{s_j f_j(\tilde{\tau}) - s_j f_j(\tau^L)}{|s_j f_j(\tau^H) - s_j f_j(\tau^L)| + \epsilon}, \quad z_j \in [-2, 2]. \tag{11}$$

Components are used only when the high-risk baseline is more conservative than the low-risk baseline. CAI averages speed, progress, and braking scores and subtracts a stability penalty from lateral deviation and jerk:

$$\mathrm{CAI} = \mathrm{mean}(S_{\mathrm{speed}}, S_{\mathrm{prog}}, S_{\mathrm{brake}}) - 0.25 P_{\mathrm{stab}}. \tag{12}$$

Positive CAI indicates that the intervention moves the low-risk output toward the matched high-risk conservative style. For models with intermediate logits, we also measure logit-level changes before trajectory decoding.

For behavior-level evaluation, the intervention rank is selected on the validation split. For each model and intervention site, we sweep $k \in \{1, 2, 4, 8\}$ and select a single diagnostic rank using validation CAI. Specifically, we apply a one-standard-error rule: after identifying the rank with the highest validation CAI, we choose the smallest k whose validation CAI falls within one standard error of that best value. This favors lower-rank interventions when their validation performance is statistically comparable to the best rank, reducing sensitivity to small validation sets. The selected k is then fixed and evaluated on the held-out test split.

V. Experimental Results

We apply RiskLens to UniAD [1], OpenDriveVLA (VLA) [52], and Diffusion Planner (DP) [2], covering modular, vision-language, and diffusion-based driving policies.

A. Experimental Setup

Risk is defined independently of the audited model using conflict measures such as distance-to-conflict, time-to-conflict, relative motion, and agent proximity. Low/high-risk labels come from dataset-specific tails, and matched pairs preserve non-risk context while enforcing a risk-score gap. Table I summarizes the audited policies. Different pair counts reflect available matched data and are not used for direct performance ranking. We use CAI as the primary behavior-level index for evaluating conservative execution; ADE is reserved only for qualitative trajectory visualization in Fig. 2.

TABLE I

EXPERIMENTAL SETUP. VLA DENOTES OPENDRIVEVLA; DP DENOTES DIFFUSION PLANNER. CAI IS USED AS THE PRIMARY BEHAVIOR-LEVEL INDEX ACROSS ALL POLICIES.

Policy	Data	Pairs	Sites	Primary Index
UniAD	nuScenes	37	BEV; tracking; planning	CAI
VLA	nuScenes	49	residual layers; tokens	CAI
DP	nuPlan	45	fusion; DiT blocks; steps	CAI

TABLE II

RISK DECODABILITY ACROSS AUDITED DRIVING POLICIES. AUROC REPORTS BINARY SAFE–RISKY CLASSIFICATION, AND ρ_s REPORTS CORRELATION WITH THE CONTINUOUS RISK SCORE.

Model	Setting	Site	AUROC ↑	ρ_s ↑
UniAD	hidden	`track dec.`	$\mathbf{0.984 \pm 0.012}$	**0.843**
	shuffled	–	0.500 ± 0.045	-0.014 ± 0.074
	random	–	0.492 ± 0.058	-0.019 ± 0.127
VLA	hidden	`L23 last`	$\mathbf{0.979 \pm 0.024}$	0.791
	metadata	ego/map/scenario	0.912 ± 0.060	**0.792**
	clean prompt	residual	0.500 ± 0.000	–
DP	strict pairs	`early DiT`	**0.62–0.64**	**0.46**
	large audit	`final denoise`	0.574 ± 0.078	0.10

B. Finding 1: Risk Information Is Decodable Across Driving Policies

We first test whether scenario-level risk is linearly decodable from hidden representations. This diagnostic only establishes whether risk information is available in latent space, not whether it is causally used. Table II reports the best probe results across models.

All three policies exhibit above-chance risk decodability, but the signal appears at different strengths and sites. UniAD shows the strongest signal at the last track-decoder state, OpenDriveVLA shows comparable late-layer decodability but is influenced by prompt-side risk proxies, and Diffusion Planner exhibits weaker, stage-dependent signals. These results motivate testing whether such decodable risk is also intervention-relevant.

C. Finding 2: DAS Reveals Actionable Risk Directions Missed by Linear Probes

Comparing linear-probe decodability with DAS interventions reveals that high accuracy does not guarantee downstream steerability. For example, UniAD's track decoder shows strong decodability (AUROC $= 0.984$) but negative held-out execution (Test CAI $= -0.135$), whereas its planning regression output yields a strong conservative response (CAI $= 0.648$). Similarly, OpenDriveVLA's highly decodable late layers (AUROC $= 0.979$) fail to execute conservatively. Effective DAS sites consistently concentrate at planning or fusion interfaces (e.g., Diffusion Planner's encoder fusion), indicating risk is causally useful only when structurally aligned with downstream prediction pathways.

D. Finding 3: Causal Risk Geometry Does Not Always Translate into Behavior-Level Execution

Finally, we evaluate if learned risk geometry induces conservative behavior on held-out pairs, reporting test CAI using validation-selected ranks. Results reveal architecture- and site-dependent bottlenecks. UniAD shows strong conservative responses at the planning regression output, and Diffusion Planner effectively transfers risk at encoder fusion but misses at early DiT blocks. Conversely, OpenDriveVLA yields near-zero test CAI, indicating that latent steerability fails to propagate to trajectory behavior. Thus, risk geometry is not uniformly executable; only specific interfaces translate internal interventions into safe actions.

TABLE III

HELD-OUT CONSERVATIVE-BEHAVIOR EXECUTION. k DENOTES THE VALIDATION-SELECTED INTERVENTION RANK.

Model	Val/Test	Site	k	Val CAI ↑	Test CAI ↑
UniAD	7/8	`reg out`	1	**0.609**	**0.648**
		`reg in`	1	0.665	0.221
		`track dec.`	1	0.247	-0.135
VLA	9/11	`L12 last`	2	**0.005**	-0.027
		`L23 all`	2	0.000	0.000
DP	9/9	`enc. fusion`	1	**0.269**	**0.345**
		`DiT block 1`	1	0.001	0.001

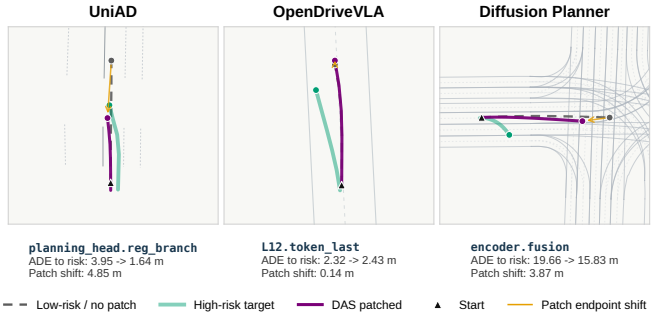

Fig. 2. Representative held-out interventions for UniAD, OpenDriveVLA, and Diffusion Planner. Each panel compares the low-risk/no-patch output, matched high-risk target, and DAS-patched output. "ADE to risk" reports trajectory ADE to the high-risk target before and after patching, and "Patch shift" denotes the endpoint displacement from the no-patch output to the patched output.

VI. CONCLUSION

Our audit reveals that end-to-end driving policies often fail to translate internal risk representations into behavioral safety. Across three architectures, risk is decodable and steerable in hidden states, but only specific fusion or planning interfaces convert it into conservative actions. Future evaluations must therefore verify not just whether models encode risk, but if it remains actionable at the decision stage.

ACKNOWLEDGMENT

This work is based upon the work supported by the National Center for Transportation Cybersecurity and Resiliency (TraCR) (a U.S. Department of Transportation National University Transportation Center) headquartered at Clemson University, Clemson, South Carolina, USA. Any opinions, findings, conclusions, and recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of TraCR, and the U.S. Government assumes no liability for the contents or use thereof.

REFERENCES

[1] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang *et al.*, "Planning-oriented autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 17 853–17 862.

[2] Y. Zheng, R. Liang, K. Zheng, J. Zheng, L. Mao, J. Li, W. Gu, R. Ai, S. E. Li, X. Zhan *et al.*, "Diffusion-based planning for autonomous driving with flexible guidance," *arXiv preprint arXiv:2501.15564*, 2025.

[3] B. Liao, S. Chen, H. Yin, B. Jiang, C. Wang, S. Yan, X. Zhang, X. Li, Y. Zhang, Q. Zhang *et al.*, "Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 12 037–12 047.

[4] P. De Haan, D. Jayaraman, and S. Levine, "Causal confusion in imitation learning," *Advances in neural information processing systems*, vol. 32, 2019.

[5] D. Chen, V. Koltun, and P. Krähenbühl, "Learning to drive from a world on rails," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 590–15 599.

[6] J. Tang, Z. Zhou, Z. He, J. Zhang, K. Zhang, and J. Pu, "Causalvad: De-confounding end-to-end autonomous driving via causal intervention," *arXiv preprint arXiv:2603.18561*, 2026.

[7] Y. Huang, Z. Qu, L. Jiang, B. Liu, and H. Zhang, "Prioritizing perception-guided self-supervision: A new paradigm for causal modeling in end-to-end autonomous driving," *arXiv preprint arXiv:2511.08214*, 2025.

[8] J. Du, Y. Song, Y. Zhao, X. Pan, J. Lian, Y. Lu, L. Wang, C. Liu, and Q. Chen, "Deconfounded lifelong learning for autonomous driving via dynamic knowledge spaces," *arXiv preprint arXiv:2603.14354*, 2026.

[9] E. Ahmadi, R. Mercurius, S. Alizadeh, K. Rezaee, and A. Rasouli, "Curb your attention: Causal attention gating for robust trajectory prediction in autonomous driving," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 779–786.

[10] C. Lei, H. Wu, Y. Ji, X. Cai, and Y. Du, "Cautraj: A causal-knowledge-guided framework for lane-changing trajectory planning of autonomous vehicles," *arXiv preprint arXiv:2512.18703*, 2025.

[11] B. Wang, H. Liao, C. Wang, B. Rao *et al.*, "Beyond patterns: Harnessing causal logic for autonomous driving trajectory prediction," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2025.

[12] Z. Zong, Y. Wang, H. Lin, L. Gan, and H. Shao, "A generalizable physics-guided causal model for trajectory prediction in autonomous driving," *arXiv preprint arXiv:2602.13936*, 2026.

[13] H. Fu, D. Zhang, Z. Zhao, J. Cui, D. Liang, C. Zhang, D. Zhang, H. Xie, B. Wang, and X. Bai, "Orion: A holistic end-to-end autonomous driving framework by vision-language instructed action generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 24 823–24 834.

[14] Y. Li, Y. Zhang, S. Thompson, E. Javanmardi, and M. Tsukada, "Causal scene narration with runtime safety supervision for vision-language-action driving," *arXiv preprint arXiv:2604.01723*, 2026.

[15] L. Giamattei, A. Guerriero, R. Pietrantuono, and S. Russo, "Causality-driven testing of autonomous driving systems," *ACM Transactions on Software Engineering and Methodology*, vol. 33, no. 3, pp. 1–35, 2024.

[16] S.-C. D. S. Generation, "Causalaf: Causal autoregressive flow for safety-critical driving scenario generation."

[17] F. Codevilla, M. Müller, A. López, V. Koltun, and A. Dosovitskiy, "End-to-end driving via conditional imitation learning," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 4693–4700.

[18] K. Chitta, A. Prakash, B. Jaeger, Z. Yu, K. Renz, and A. Geiger, "Transfuser: Imitation with transformer-based sensor fusion for autonomous driving," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 11, pp. 12 878–12 895, 2023.

[19] B. Jiang, S. Chen, Q. Xu, B. Liao, J. Chen, H. Zhou, Q. Zhang, W. Liu, C. Huang, and X. Wang, "Vad: Vectorized scene representation for efficient autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8340–8350.

[20] M. Pourkeshavarz, J. Zhang, and A. Rasouli, "Cadet: a causal disentanglement approach for robust trajectory prediction in autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 874–14 884.

[21] J. Sun, H. Zhang, G. Ding, and Z. Zheng, "Echo planning for autonomous driving: From current observations to future trajectories and back," *arXiv preprint arXiv:2505.18945*, 2025.

[22] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," *arXiv preprint arXiv:1907.02893*, 2019.

[23] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio, "Toward causal representation learning," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 612–634, 2021.

[24] Y. Tian, K. Pei, S. Jana, and B. Ray, "Deeptest: Automated testing of deep-neural-network-driven autonomous cars," in *Proceedings of the 40th International Conference on Software Engineering*, 2018, pp. 303–314.

[25] M. Zhang, Y. Zhang, L. Zhang, C. Liu, and S. Khurshid, "Deeproad: Gan-based metamorphic testing and input validation framework for autonomous driving systems," in *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering*, 2018, pp. 132–142.

[26] D. J. Fremont, T. Dreossi, S. Ghosh, X. Yue, A. L. Sangiovanni-Vincentelli, and S. A. Seshia, "Scenic: A language for scenario specification and scene generation," in *Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation*, 2019, pp. 63–78.

[27] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Proceedings of the 1st Annual Conference on Robot Learning*, 2017, pp. 1–16.

[28] M. Althoff, M. Koschi, and S. Manzinger, "Commonroad: Composable benchmarks for motion planning on roads," in *2017 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2017, pp. 719–726.

[29] C. Xu, W. Ding, W. Wang, Y. Cao, X. Chen, S. Li, Z. Ding, Z. Liu, Z. Tu, J. Guo *et al.*, "Safebench: A benchmarking platform for safety evaluation of autonomous vehicles," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 25 667–25 682.

[30] R. P. McDonald, "Judea pearl. causality: Models, reasoning, and inference. cambridge: Cambridge university press. 384 pp., 2000, isbn 0521773628." *Psychometrika*, vol. 67, no. 2, pp. 321–322, 2002.

[31] J. Peters, D. Janzing, and B. Schölkopf, *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 2017.

[32] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[33] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.

[34] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, "Network dissection: Quantifying interpretability of deep visual representations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6541–6549.

[35] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. Sayres, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)," in *Proceedings of the 35th International Conference on Machine Learning*, 2018, pp. 2668–2677.

[36] K. Meng, D. Bau, A. Andonian, and Y. Belinkov, "Locating and editing factual associations in gpt," *Advances in neural information processing systems*, vol. 35, pp. 17 359–17 372, 2022.

[37] J. Vig, S. Gehrmann, Y. Belinkov, S. Qian, D. Nevo, Y. Singer, and S. Shieber, "Causal mediation analysis for interpreting neural nlp: The case of gender bias," in *Advances in Neural Information Processing Systems*, 2020.

[38] C. Olah, N. Cammarata, L. Schubert, G. Goh, M. Petrov, and S. Carter, "Zoom in: An introduction to circuits," *Distill*, 2020.

[39] A. Geiger, Z. Wu, C. Potts, T. Icard, and N. Goodman, "Finding alignments between interpretable causal variables and distributed neural representations," in *Causal Learning and Reasoning*. PMLR, 2024, pp. 160–187.

[40] A. Geiger, D. Ibeling, A. Zur, M. Chaudhary, S. Chauhan, J. Huang, A. Arora, Z. Wu, N. Goodman, C. Potts *et al.*, "Causal abstraction: A theoretical foundation for mechanistic interpretability," *Journal of Machine Learning Research*, vol. 26, no. 83, pp. 1–64, 2025.

[41] T. Trim and T. Grayston, "Mechanistic interpretability of reinforcement learning agents," *arXiv preprint arXiv:2411.00867*, 2024.

[42] S. Basu, S. Patel, P. Sheth, B. Muralidharan, N. Elamaran, A. Kinra, and R. Batniji, "Mechanistic interpretability of reinforcement learning

in medicaid care coordination," *BMJ Health & Care Informatics*, vol. 33, no. 1, p. e101935, 2026.

[43] J. H. Lee, A. Lauscher, and S. V. Albrecht, "Towards ethical multi-agent systems of large language models: A mechanistic interpretability perspective," *arXiv preprint arXiv:2512.04691*, 2025.

[44] Y. Shi, C. Li, Y. Wang, Y. Zhao, A. Pang, S. Yang, J. Yu, and K. Ren, "Dissecting and mitigating diffusion bias via mechanistic interpretability," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 8192–8202.

[45] B. Tinaz, Z. Fabian, and M. Soltanolkotabi, "Emergence and evolution of interpretable concepts in diffusion models," 2025.

[46] A. Helbling, T. H. S. Meral, B. Hoover, P. Yanardag, and D. H. Chau, "Conceptattention: Diffusion transformers learn highly interpretable features," in *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 267, 2025. [Online]. Available: https://arxiv.org/abs/2502.04320

[47] J. C. Hayward, "Near-miss determination through use of a scale of danger," *Highway Research Record*, vol. 384, pp. 24–34, 1972.

[48] R. Krajewski, J. Bock, L. Kloeker, and L. Eckstein, "The highd dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2018, pp. 2118–2125.

[49] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 621–11 631.

[50] H. Caesar, J. Kabzan, K. S. Tan, W. K. Fong, E. Wolff, A. Lang, L. Fletcher, O. Beijbom, and S. Omari, "nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles," *arXiv preprint arXiv:2106.11810*, 2021.

[51] A. Geiger, H. Lu, T. Icard, and C. Potts, "Causal abstractions of neural networks," *Advances in neural information processing systems*, vol. 34, pp. 9574–9586, 2021.

[52] X. Zhou, X. Han, F. Yang, Y. Ma, V. Tresp, and A. Knoll, "Opendrivevla: Towards end-to-end autonomous driving with large vision language action model," 2025. [Online]. Available: https://arxiv.org/abs/2503.23463