What Do They See? Interpreting Complex Road Scenarios Through the Eyes of Vision-Language-Action Models for Safe and Trustworthy Autonomous Vehicle Learning

Kalpana Panda[1,2] **Wesley Maia**[2] **Vinti Agarwal**[1] **Ross Greer**[2]

[1]Department of Computer Science and Information Systems,
Birla Institute of Technology and Science, Pilani, India
[2]Machine Intelligence, Interaction, and Imagination (Mi[3]) Lab,
University of California, Merced, United States

arXiv:2607.16938v1 [cs.CV] 18 Jul 2026

Abstract: End-to-end autonomous driving models are now able to navigate complex road scenarios, mapping raw sensor observations directly to observed paths for open-loop evaluation and often effective driving in closed-loop evaluation. Yet the internal logic of these safety-critical systems remains largely opaque, due to the complexity of traffic scenes. We propose a counterfactual ablation framework called Counterfactual Vision Action Analysis (CVAA) that systematically removes individual detected objects from front-camera images using photorealistic generative inpainting to prepare counterfactual sets to evaluate the difference in the model's response. This isolates the causal effect of each object's presence on the model's planning behaviour. Applied to the Alpamayo 1 trajectory predictor across 210 nuScenes driving scenes, we create a dataset Counter -nuScenes, using which we see that vehicles and pedestrians within the model's 'path' dominate causal influence as expected, while traffic lights, as expected, exert disproportionate effect relative to their image footprint. However, we also find cases where the model responds strongly to objects a human driver would consider irrelevant. This brings forth a deeper question: does the model itself view the scene as a sum of individual objects influencing the outcome, or does it encode an entirely different set of internal features that do not correspond to human-legible scene elements? To further understand this, we compare intermediate representations of original and inpainted image pairs using mechanistic interpretability techniques and examine the effect of the removal through the various model layers. Together, these two stages offer a path from behavioral auditing to representational understanding, creating explainable driving systems and solidifying human-AI trust.

Keywords: multimodal interpretability, measurable object importance, scene understanding

1 Introduction

Autonomous driving systems must interpret complex scenes containing interacting vehicles, pedestrians, cyclists, traffic signals, road infrastructure, and work-zone elements. Large-scale datasets such as KITTI [1], nuScenes [2], the Waymo Open Dataset [3], and BDD100K [4] have driven major progress in perception, tracking, forecasting, and planning. More recently, autonomous-driving research has moved toward unified models that connect visual scene understanding with language reasoning and action generation, including UniAD [5], DriveGPT4 [6], LMDrive [7], DriveLM [8], DriveVLM [9], and EMMA [10].

Despite this progress, it remains difficult to determine which scene objects actually influence a model's driving decision. A model may generate a plausible trajectory or a convincing language rationale without being causally driven by the correct visual evidence. This is especially concerning in safety-critical settings: a driving model should respond to a lead vehicle, crossing pedestrian, or traffic light, while remaining relatively insensitive to irrelevant background objects. We therefore ask an object-level causal question: which visible objects drive a vision-based model's trajectory prediction?

Existing causal and perturbation-based evaluations in driving typically operate on structured state representations, simulator abstractions, or agent trajectories [11, 12, 13]. Visual attribution methods such as saliency maps, Grad-CAM, RISE, and SHAP [14, 15, 16, 17] instead operate on images, but pixel-level explanations can be hard to audit in dense traffic scenes and may conflate object importance with artifacts introduced by position, masking, or occlusion [18, 19, 20]. We address this gap by intervening directly on semantic objects in the camera image while keeping the edited scene visually plausible.

We introduce *Counterfactual Vision Action Analysis (CVAA)*, a counterfactual object-attribution framework, and *Counter-nuScenes*, a benchmark dataset for vision-based autonomous-driving models. Given a front-camera traffic scene, we detect and segment individual objects, remove each one using photorealistic generative inpainting, rerun trajectory inference, and measure the resulting shift relative to the model's original trajectory distribution. By replacing removed objects with contextually plausible background rather than black boxes, blur, or simple occlusion, the framework aims to isolate the causal effect of each object's presence while reducing distribution shift from unrealistic perturbations [21, 22, 23, 24].

We evaluate this framework on nuScenes front-camera scenes using the Alpamayo 1 trajectory predictor. For each scene, we construct a counterfactual semantic set in which each detected object is removed independently, and we rank objects by trajectory-shift metrics computed against the model's own original prediction. Unlike minADE and minFDE, which measure accuracy against a singular "ground truth" instantiation of all possible driving behaviors [25], our metrics measure behavioral sensitivity: the goal is not to evaluate whether the model predicts the human driver correctly, but to identify which objects causally affect its output. Our trajectory-aimed approach removes the additional layer of oversimplification of a text output given, which may not register changes in trajectory as long as the broad path of the vehicle remains the same. This helps to further enhance the safety, as even slight trajectory changes can prove to be injurious in certain scenes. This is especially important in the context of safety, as it is not the verbal explanation which determines the safety of the scene, but rather the output trajectory which dictates vehicle motion.

Beyond black-box behavioral attribution, we also use the resulting high-influence objects as entry points for white-box analysis. Comparing original and inpainted representations, and applying mechanistic interpretability tools such as sparse autoencoders and activation patching [26, 27, 28, 29], we analyze whether behaviorally important objects correspond to recoverable internal features or causal model components.

Contributions. (C1) We introduce **Counter-nuScenes**, a counterfactual benchmark built from 210 nuScenes scenes with photorealistic per-object inpainting, producing 3,062 in-distribution counterfactual image pairs. **(C2)** We propose **CVAA AD** and **CVAA FD**, two seed-stable distributional shift metrics that quantify each object's causal influence on predicted trajectories without requiring ground-truth poses. **(C3)** We provide a **black-box evaluation** of Alpamayo 1, systematically measuring attribution patterns across object class, mask size, and scene structure. **(C4)** We conduct a **white-box mechanistic analysis** that traces layer-wise causal propagation, identifies four propagation regimes, and reveals the trajectory expert as a nonlinear amplifier of object-level perturbations.

2 Related Work

Unified driving models. Autonomous-driving research has shifted from modular perception–prediction–planning stacks toward models that couple scene understanding, reasoning, and action. Planning-oriented systems such as UniAD [5] optimize intermediate perception and forecasting for downstream planning, while driving VLM/VLA systems such as DriveGPT4 [6], LMDrive [7], DriveLM [8], DriveVLM [9], OpenDriveVLA [30], AutoVLA [31], Reasoning-VLA [32], and Alpamayo 1 [33, 34] combine visual context with language or action generation. These systems improve inspectability, but generated rationales do not by themselves establish which visual entities causally influenced the final trajectory.

Causal evaluation and visual attribution. Causal confusion, ChauffeurNet, and CausalAgents show that learned driving policies and forecasters can depend on non-causal correlates or can be evaluated through targeted perturbations [11, 12, 13]. In parallel, saliency, Grad-CAM, meaningful perturbations, RISE, and SHAP provide image-level attribution tools [14, 15, 35, 16, 17]. Object-level methods such as PixelSHAP move closer to traffic-scene auditing [36], but artificial masks or occlusions can introduce out-of-distribution evidence. Our work combines the causal question from driving perturbation studies with object-level interventions performed directly on camera images.

Counterfactual editing and mechanistic analysis. Generative counterfactual explanations and inpainting-based editing reduce perturbation artifacts by replacing removed content with plausible background [21, 22]. Modern tools such as SAM2 [24] and LaMa [23] make object-level counterfactual image sets practical at scale. Behavioral attribution, however, still leaves open how influential objects are represented internally. Mechanistic interpretability methods, including causal tracing, activation patching, and sparse autoencoders [26, 27, 28, 29], and recent multimodal analyses such as NOTICE and VLM-SAE work [37, 38], provide tools for connecting object-level counterfactual effects to internal model features.

3 Methodology

Our framework, *Counterfactual Vision Action Analysis (CVAA)*, probes the causal influence of individual scene objects on the trajectory predictions of a VLA-based autonomous driving model by constructing a controlled counterfactual benchmark and measuring the resulting distributional shift in predicted futures. We describe each component in turn.

3.1 Benchmark Construction

Base dataset. Our Counter-nuScenes benchmark is built on the nuScenes v1.0-trainval split [2], which comprises 700 driving scenes of approximately 20 seconds each, recorded in Boston and Singapore across a range of urban traffic conditions. Each scene provides six temporally synchronized camera streams, LiDAR sweeps, and high-frequency ego-pose measurements from GPS/IMU at up to 20 Hz.

Frame selection. We restrict our evaluation to front-camera (`CAM_FRONT`) keyframes, which focus on the model's primary visual input and support a more localized interpretation of counterfactual effects. To ensure sufficient ego-motion context, we retain only frames preceded by at least 1.6 seconds of continuous pose data, matching the history window expected by the model. Within each scene, we detect the number of visible agents using YOLOv8 [39] and select the single keyframe with the highest object count, yielding one maximally informative observation per scene. From the resulting pool, after cleaning the data, we sample **210 frames**, spanning scenes with 6 to 32 valid detected objects and covering a wide range of traffic densities.

Object segmentation. For each selected frame, we apply the Segment Anything Model 2 (SAM2) [24] to obtain per-object instance masks. SAM2 is prompted with the bounding boxes produced by YOLOv8, and its masks are refined at the pixel level to avoid bleeding across object boundaries.

3.2 Counterfactual Semantic Sets

For the purpose of CVAA and given the instance masks for a scene, we construct a *counterfactual semantic set Counter-nuScenes*: a collection of images in which each object has been individually removed via photorealistic inpainting while all other scene content remains unchanged. This produces a paired corpus of one original image and N counterfactual images per scene, where N is the number of detected objects. **Inpainting approach.** Object removal is performed using Gemini's

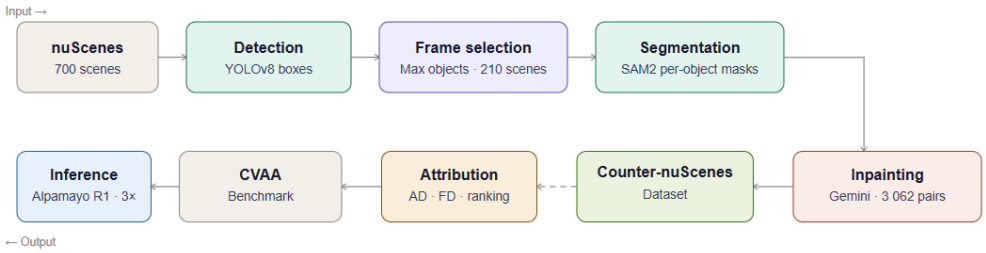

Figure 1: Counter-nuScenes and CVAA construction pipeline.

imgen inpainting tool [40]. We also maintain an open-source ablation pipeline based on LaMa [23] combined with FLUX Fill 1.0 [41]. We deliberately use photorealistic inpainting rather than simpler alternatives such as color filling, Gaussian blurring, or copy-paste occlusion. This design choice distinguishes our setting from prior attribution methods in autonomous driving, where masks were blacked out [36] or semantic image pairs were used as surrogates [37]. Photorealistic inpainting keeps counterfactual images closer to the model's data distribution: the intended change is confined to the masked region and corresponds to the semantic absence of the object, rather than to distribution shift introduced by an unrealistic perturbation. Consequently, changes in model output can be interpreted as responses to the removed object rather than to rendering artifacts.

3.3 Attribution Metrics

For each scene, we run the model on the original image to obtain a reference trajectory distribution and on each counterfactual image to obtain a variant distribution. We quantify the causal influence of each removed object by measuring the displacement between these two distributions.

Trajectory representation. The model produces K stochastic trajectory samples $\{\mathbf{p}_k\}_{k=1}^{K}$, where each $\mathbf{p}_k$ is a sequence of T future ego-centric waypoints. We project trajectories onto the XY plane and summarize each predictive distribution by its mean trajectory, which is more stable across random seeds than selecting a single sample.

Average Deviation (AD) and Final Deviation (FD). Let $\bar{\mathbf{p}}^{\text{orig}}$ and $\bar{\mathbf{p}}^{\text{var}}$ denote the mean trajectories of the original and counterfactual runs, respectively. We define:

$$\begin{aligned} \bar{\mathbf{p}} &= \frac{1}{K}\sum_{k=1}^{K}\mathbf{p}_k \in \mathbb{R}^{T\times 2}, \\ \mathrm{AD} = \frac{1}{T}\sum_{t=1}^{T}\left\|\bar{\mathbf{p}}_t^{\text{var}} - \bar{\mathbf{p}}_t^{\text{orig}}\right\|_2, &\quad \mathrm{FD} = \left\|\bar{\mathbf{p}}_T^{\text{var}} - \bar{\mathbf{p}}_T^{\text{orig}}\right\|_2. \end{aligned} \tag{1}$$

AD captures the *integrated* influence of an object over the full prediction horizon, while FD isolates its effect on the *terminal position*, a quantity directly relevant to planning safety margins. Together, they provide complementary views: an object may have a large AD, indicating early influence that later decays, but a modest FD, or vice versa.

Object-level attribution score. For a scene with N objects, removing object i yields the pair $(\mathrm{AD}_i, \mathrm{FD}_i)$. We rank objects by AD as our primary attribution score, breaking ties by FD. This

ranking constitutes our *counterfactual object attribution* (COA) ordering for the scene. Aggregating COA rankings across all 210 scenes allows us to analyze which object categories are systematically most influential to the model's trajectory predictions, and to compare the structure of these attributions between the black-box and white-box evaluation regimes (Section 4).

Relationship to existing metrics. AD and FD are conceptually related to the standard minADE and minFDE metrics used in trajectory forecasting [2], but serve a fundamentally different purpose. minADE and minFDE measure *accuracy* relative to a ground-truth future, selecting the best candidate trajectory. AD and FD instead measure *distributional shift* relative to a reference prediction, using the full candidate set. Using a min-over-K estimator here would be inappropriate: it would reward any variant that happens to produce one trajectory close to the original distribution by chance, conflating sampling noise with causal influence. The mean-based estimator in Equation 1 avoids this confound.

4 Experimental Setup

To investigate Alpamayo's behavior, we evaluate the model under two complementary regimes. First, in the **black-box** setting, we treat Alpamayo 1 as an opaque function and measure object influence solely through AD and FD on *CVAA* (Section 3.3). We run Alpamayo 1 three times on the dataset using seeds 42, 7, and 123. During these runs, inputs from all cameras except the front camera are zeroed out, and the model receives a single still image for each original or inpainted scenario. This isolates the effect of removing one object at a time from the front-camera view. We also create a dashboard A.2 that gives a visual reference to compare the importance of different objects in a scene.

Second, in the **white-box** setting, we use the *Counter-nuScenes* image pairs to inspect internal model behavior. Alpamayo 1 comprises a vision encoder (27 blocks, a VLM (36 layers, 3006-token sequence with 180 tokens per camera image after spatial merging), and a trajectory expert (36 denoising layers).

For each removed object, its bounding box is projected onto the model's patch grids to identify the corresponding token positions $\mathcal{T}$ in both the vision encoder and VLM. We compute layer-wise cosine deltas between original and inpainted hidden states at object, visual, and global token positions across all three modules [42, 43, 44], yielding a per-layer attribution signal that localises where in the model the object's removal registers. We additionally apply a Logit Lens [45] to track how the object's class representation evolves across layers, compute spatial KL divergence maps over visual patch positions [44], and extract attention weights from the trajectory handoff token to the object's patch positions. Delta curves are computed for all layers; Logit Lens and attention analyses are run at representative layers in each module.

5 Results

5.1 Black-Box Evaluation

Per-Class Deviation. Table 3 reports mean AD and FD per object class. `Class bus` seems to produce the highest deviation (AD $= X, Y$ m), while `class umbrella` objects appear to have negligible impact. However, this is additionally affected by their lower numbers, resulting in relatively skewed AD/FD statistics, while more popular object classes have their influence more smoothed out. One notable point, however, is the influence that traffic lights appear to possess, outranking both cars and people, which is analogous to their importance on the road.

Rank Stability. Since raw AD and FD statistics are scene-specific, we proceed to evaluate objects on the basis of their ranking within the scenes, based on AD and FD. On checking the rank stability between runs with different scenes, an interesting point to note is that objects rarely maintain their rank. As visible in 1, only 5.6% of objects keep their rank with respect to AD, and 6.4% with respect to FD, with 14.6% and 16% of objects in a top 3 rank remaining within the top 3, respectively, with

Table 1: **Rank Stability Summaries (3 Seeds).**

Rank Stability Summary (AD)	
Top-3 Objects	
Exact rank kept	14.6%
Remain in top-3	39.7%
All Objects	
Exact rank kept	5.7%
Rank shift $\leq \pm 2$	34.1%

Rank Stability Summary (FD)	
Top-3 Objects	
Exact rank kept	16.0%
Remain in top-3	39.5%
All Objects	
Exact rank kept	6.4%
Rank shift $\leq \pm 2$	35.2%

the average rank range being 4.691 for AD and 4.641 for FD. This strongly indicates that object importance is not stable and hence that the model rarely takes cues from a single object.

Scene-Level Deviation. Based on rankings derived by averaging AD and FD for each object across the three runs, we also examine scenes individually.

Per-scene distribution of importance. Figure 4 shows the per-scene z-score of the maximum-AD object, where $z = (\max(\text{AD}) - \bar{\text{AD}})/\sigma_{\text{AD}}$ measures how many standard deviations the most influential object lies above the scene mean. The distribution spans $z \in [1.0,\ 4.52]$ with mean 2.43, median 2.35, and IQR $[1.91,\ 2.96]$, indicating that in the typical scene some object exerts a disproportionate influence relative to its peers. 148 of 210 scenes (70.5%) exceed the conventional $z>2$ threshold, and 48 scenes (22.9%) exceed $z>3$. FD z-scores are closely aligned (mean 2.41, median 2.34, $z>2$: 148, $z>3$: 43), with no systematic decoupling between integrated and terminal deviation.

The right-skewed tail reflects the structural heterogeneity of urban driving scenarios: dense junction scenes distribute influence across many agents and suppress z, while sparse environments (where a single isolated object is the dominant scene cue) drive z into the extreme tail. This dominance is best understood as a property of the scene's deviation landscape rather than a stable attribution to one specific object, as rank orderings show considerable seed-to-seed variance (Table 1). This suggests that the model is sensitive to the presence of a high-influence region rather than to one uniquely identifiable agent, which we discuss further in Section 8 . **Object count vs peak deviation per scene.** Crucially, object count shows *no relationship* with peak deviation (Spearman $\rho = 0.11$ for AD,+ 0.12 for FD), suggesting that object presence in certain high-priority regions matters more than scene density in driving trajectory shifts.

Mask Size. Mask area correlates positively with both AD and FD (Pearson $r = +0.25$, Spearman $\rho = +0.22$), indicating that physically larger objects exert greater causal influence on the model's predicted trajectory when removed. This is further reflected in the rank structure: rank-1 objects (highest AD) average 38,706 pixels, falling sharply to 13,932 at rank 2 and below 5,000 beyond rank 5 (Spearman $\rho = -0.56$, $p \ll 0.001$), suggesting that object scale is a primary structural predictor of deviation within a scene.

Traffic Light Colour. Of 531 traffic light objects, pixel-level colour classification detected 119 (22.4%); the remainder lacked a clearly dominant red or green channel, likely due to grey housing, night conditions, or overexposure. Among detected lights, red-state lights produce substantially higher deviation than green-state (mean AD 0.758 vs 0.229 m; mean FD 2.241 vs 0.672 m), consistent with the intuition that a stationary ego at a red light has a more constrained and predictable trajectory. Thus, removing an object in that context causes a larger distributional shift than when the ego is in motion at a green light.

Distributional Shift vs. Ground-Truth Accuracy. Of 2,975 objects, 1,468 (49.3%) exhibit a negative Δ minADE , meaning their removal causes the model's best trajectory to move *closer* to ground truth despite a distributional shift. Among the 360 high-AD objects (top 25% by AD, threshold 0.516 m), the Spearman correlation between AD and Δ minADE is near zero ($\rho = +0.03$), confirming that the magnitude of distributional shift carries no consistent signal about whether accuracy improves or degrades. The most extreme case is scene-0708, where two cars produce Δ minADE $= -2.49$ and -1.90 m respectively, suggesting the model was substantially over-relying on those objects.

Per-Class Rank Distribution. Table 2 shows how often each class occupies the top AD rank across 210 scenes. Cars rank first most frequently (78 scenes, 37.1%), followed by persons (48, 22.9%) and traffic lights (41, 19.5%). However, when normalised by class count, buses and trucks rank first at a disproportionately high rate (19.1% and 16.4% of their appearances respectively), consistent with their larger physical scale and correspondingly larger masks.

The interesting observation, however, comes in the form of certain seemingly ranking objects having disproportionately high rankings. In 1.0% of scenes, a stop sign ranks above all other agents, and in one scene a tennis racket is the single most influential object. Even within plausible classes, the ranking frequently defies intuition: a pedestrian who has already crossed the road above one stepping into the ego's path, a ball-shaped fixture on the side of the road above someone crossing the road, a traffic light whose signal isn't visible from the photograph angle over a car immediately ahead of the camera. These cases seem to be the rule, not the exception. 49.3% of objects improve ground-truth accuracy when removed suggesting the model is not simply tracking the most safety-critical agent, but looking at the image a lot more holistically than we may perceive. Additionally, it points to the fact that there are multiple objects that the model draws consolidated cues from (a person on the crosswalk and 2 different red light signals in front of the car all point to the fact that the vehicle should stop). Combined with the low rank stability across seeds (Section 1), this points to a fundamental limitation of single-object counterfactual attribution: the model's response to any one removal is entangled with its broader scene representation.

Table 2: Per-class rank-1 frequency across 210 scenes. Classes never ranking first are omitted.

Class	N	#1	%1	#Top 3	%1/Total
Car	1566	78	37.1	272	5.0
Person	604	48	22.9	134	7.9
Traffic light	531	41	19.5	122	7.7
Truck	183	30	14.3	61	16.4
Bus	47	9	4.3	18	19.1
Stop sign	8	2	1.0	3	25.0
Tennis racket	1	1	0.5	1	100.0
Bicycle	28	1	0.5	7	3.6

5.2 White-Box Evaluation

We structure our analysis layer-by-layer, tracing how the removal of each object propagates through Alpamayo 1's three modules: vision encoder, language model, and trajectory expert.

Visual Encoding and LM Propagation. The vision encoder delta at the removed object's patch positions is on the order of 10^{-7} and negative across all categories (Figure 3, bottom right), indicating the inpainted region is more self-consistent in feature space than the original, a consequence of photorealistic inpainting. Despite this, the signal does not transfer cleanly into the LM: the correlation between late vision encoder delta and early LM delta is weak or negative for all classes (Pearson $r \in [-0.41, +0.13]$), with a sign reversal for car (-5.94) and other (-4.81), suggesting the two representations diverge differently at the merge step. Additionally, curves are identical across categories at this stage, indicating that the vision encoder does not differentiate between object classes.

Within the LM, all signals are flat for layers 0–15. From layer 15, curves diverge sharply by category: bus reaches ≈ 0.06 at layer 35 in the object-token panel, truck ≈ 0.02, while car and bicycle remain near zero throughout (Figure 3, top left). Comparing the object-token and all-visual delta panels shows the perturbation is spatially contained for most classes; bus is the exception, where the all-visual delta substantially exceeds the object-token delta, indicating the model re-interprets surrounding context when a large occluding vehicle is removed.

Trajectory Handoff and Four Propagation Regimes. The global delta (last token, Figure 3, bottom left) peaks at layers 20–25 before decaying, with bus and truck peaking highest. This fall in the curve indicates that, despite recognition of an extra object in the original case, it resolves to a stabilised state (the difference between the two is suppressed). This explains the globally small

handoff delta (median $\delta_{\text{traj}} = 0.0002$, confirming that most object absences barely register at the VLM–expert interface.

Classifying variants by whether δ_{traj} and $|\Delta \text{minADE}|$ are each above or below their medians reveals four regimes. *Coupled* variants (approx. 33.0%) propagate cleanly from LM to output; most prevalent in bus (62.9%) and motorcycle (66.7%). *Transparent* variants (approx. 33.1%) produce negligible signal at every stage; dominant in car (36.2%). *Decoupled* variants (approx. 16.9%) register at the handoff, often via a late spike through the LM, but the both seem to converge; the model noticed two different scenes acted the same despite it. *Silent* variants (approx. 16.9%) show negligible handoff delta yet produce substantial output AD (up to 1.18 in scene-0293), indicating influence that bypasses the trajectory token entirely.

Expert Amplification. Across all categories and 100% of variants, the trajectory expert reacts in a way that causes divergence based on what it received from the VLM (mean 115–$180\times$, median 47–$106\times$). Amplification is most extreme for silent variants, where $\delta_{\text{traj}} \approx 0.0001$ yet ratios exceed $1000\times$ in scenes 0293, 0464, and 0637. The expert is therefore not a passive decoder: it is a sensitive nonlinear amplifier that resolves near-zero handoff perturbations into large distributional shifts, and is the primary mechanism by which silent variants produce high output deviation despite leaving no trace at the trajectory token.

6 Discussion

The black-box and white-box results are broadly consistent but reveal a more complex picture than either alone would suggest. Individual object rankings are unstable, indicating that class-level patterns are reliable while object-level attribution is not. The four propagation regimes show that the trajectory handoff token is not a universal bottleneck: silent variants exert causal influence through a pathway that bypasses it entirely, while decoupled variants confirm the expert actively suppresses differences it receives rather than faithfully propagating them. The LM's late-layer convergence between the two runs, the global delta peaking then decaying before handoff is consistent with a strong scene prior that stabilises the representation regardless of which specific objects are present, which may explain the model's robustness to transparent object removals in dense scenes. The Logit Lens drop ($P_{\text{orig}} - P_{\text{inp}}$, $\sim 10^{-5}$) is not only near-zero but incoherent: classes such as bus assign *higher* probability to their own label in the inpainted run than in the original, with no consistent relationship to the object's actual presence. This indicates the model does not perceive object identity as a vocabulary-level concept at any layer: the visual tokens encode scene geometry and relational context, not categorical object labels, and standard language model interpretability tools do not transfer to this representational regime.

7 Limitations

Our evaluation is limited to a single model (Alpamayo 1) and the nuScenes front-camera distribution; generalisation to other VLAs and datasets remains unvalidated. More fundamentally, the white-box analysis exposes the absence of interpretability tools designed for VLA architectures. Cosine delta curves and Logit Lens, both adapted from language model research, measure what changes between two forward passes but cannot explain *why*, nor identify which internal computations are causally responsible for the output difference. The silent quadrant in particular, where large output deviation occurs without any signature at the trajectory token, points to influence pathways that existing probes cannot resolve. Developing probes that operate in the continuous geometry of multimodal hidden states, rather than through vocabulary projection, is a prerequisite for deeper mechanistic understanding of VLA decision-making.

8 Conclusion

We introduced CVAA and Counter-nuScenes, a counterfactual model-agnostic benchmark and dataset respectively that measures the causal influence of individual scene objects on VLA trajectory predictions via photorealistic inpainting and distributional deviation metrics (AD and FD). Applied to Alpamayo 1, the pipeline surfaces consistent class-level sensitivities: large vehicles cause the most deviation, mask size and lateral position are structural predictors. White-box analysis, on the other hand, reveals that this influence operates through multiple internal pathways, not all of which are visible at the trajectory handoff. Our work thus opens a window of insight into how autonomous vehicles 'see' scenes as more than the sum of individual objects, where no one part can heavily disproportionately impact the models' outputs. Future work includes applying it to other VLAs, evaluating text-similarity in VLA prompted outputs, incorporating temporal multi-frame counterfactuals, and developing interpretability probes suited to scene-level relational representations.

References

[1] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012. doi:10.1109/CVPR.2012.6248074.

[2] 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 *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11618–11628, 2020. doi:10.1109/CVPR42600.2020.01164.

[3] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofeev, S. Ettinger, M. Krivokon, A. Gao, A. Joshi, S. Zhao, S. Cheng, Y. Zhang, J. Shlens, Z. Chen, and D. Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. *arXiv*, 2019.

[4] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2636–2645, 2020.

[5] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang, L. Lu, X. Jia, Q. Liu, J. Dai, Y. Qiao, and H. Li. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17853–17862, 2023.

[6] Z. Xu, Y. Zhang, E. Xie, Z. Zhao, Y. Guo, K.-Y. K. Wong, Z. Li, and H. Zhao. DriveGPT4: Interpretable end-to-end autonomous driving via large language model, 2023.

[7] H. Shao, Y. Hu, L. Wang, S. L. Waslander, Y. Liu, and H. Li. LMDrive: Closed-loop end-to-end driving with large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[8] C. Sima, K. Renz, K. Chitta, L. Chen, H. Zhang, C. Xie, J. Beisswenger, P. Luo, A. Geiger, and H. Li. DriveLM: Driving with graph visual question answering, 2024.

[9] X. Tian, J. Gu, B. Li, Y. Liu, Y. Wang, Z. Zhao, W. Zhan, P. Wang, H. Wang, K. K. Zhao, et al. DriveVLM: The convergence of autonomous driving and large vision-language models, 2024.

[10] Waymo Research. EMMA: End-to-end multimodal model for autonomous driving, 2024.

[11] P. de Haan, D. Jayaraman, and S. Levine. Causal confusion in imitation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[12] M. Bansal, A. Krizhevsky, and A. Ogale. ChauffeurNet: Learning to drive by imitating the best and synthesizing the worst, 2019.

[13] R. Roelofs, L. Sun, B. Caine, K. S. Refaat, B. Sapp, S. Ettinger, and W. Chai. CausalAgents: A robustness benchmark for motion forecasting using causal relationships. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

[14] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014. URL `https://arxiv.org/abs/1312.6034`.

[15] 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 *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. doi:10.1109/ICCV.2017.74.

[16] V. Petsiuk, A. Das, and K. Saenko. RISE: Randomized input sampling for explanation of black-box models. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

[17] S. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions, 2017. URL https://arxiv.org/abs/1705.07874.

[18] R. Greer, J. Isa, N. Deo, A. Rangesh, and M. M. Trivedi. On salience-sensitive sign classification in autonomous vehicle path planning: Experimental explorations with a novel dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 636–644, 2022.

[19] R. Greer, A. Gopalkrishnan, J. Landgren, L. Rakla, A. Gopalan, and M. Trivedi. Robust traffic light detection using salience-sensitive loss: Computational framework and evaluations. In *2023 IEEE Intelligent Vehicles Symposium (IV)*, pages 1–7. IEEE, 2023.

[20] R. Greer, A. Gopalkrishnan, N. Deo, A. Rangesh, and M. Trivedi. Salient sign detection in safe autonomous driving: Ai which reasons over full visual context. In *27th International Technical Conference on the Enhanced Safety of Vehicles (ESV) National Highway Traffic Safety Administration*, number 23-0333, 2023.

[21] C.-H. Chang, E. Creager, A. Goldenberg, and D. Duvenaud. Explaining image classifiers by counterfactual generation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

[22] R. Zhang, W. Li, P. Wang, C. Guan, J. Fang, Y. Song, J. Yu, B. Chen, W. Xu, and R. Yang. AutoRemover: Automatic object removal for autonomous driving videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12853–12861, 2020. doi: 10.1609/aaai.v34i07.6982.

[23] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, and V. Lempitsky. Resolution-robust large mask inpainting with fourier convolutions, 2021. URL https://arxiv.org/abs/2109.07161.

[24] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer. Sam 2: Segment anything in images and videos, 2024. URL https://arxiv.org/abs/2408.00714.

[25] R. Greer, N. Deo, and M. Trivedi. Trajectory prediction in autonomous driving with a lane heading auxiliary loss. *IEEE Robotics and Automation Letters*, 6(3):4907–4914, 2021.

[26] N. Elhage, N. Nanda, C. Olsson, T. Henighan, N. Joseph, B. Mann, A. Askell, Y. Bai, A. Chen, T. Conerly, N. DasSarma, D. Drain, D. Ganguli, Z. Hatfield-Dodds, D. Hernandez, S. Johnston, A. Jones, J. Kernion, L. Lovitt, K. Ndousse, D. Amodei, T. Brown, J. Clark, J. Kaplan, S. McCandlish, and C. Olah. A mathematical framework for transformer circuits. https://transformer-circuits.pub/2021/framework/index.html, 2021.

[27] K. Meng, D. Bau, A. Andonian, and Y. Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[28] T. Bricken, A. Templeton, J. Batson, B. Chen, A. Jermyn, T. Conerly, N. L. Turner, C. Anil, C. Denison, A. Askell, R. Lasenby, Y. Wu, S. Kravec, N. Schiefer, T. Maxwell, N. Joseph, Z. Hatfield-Dodds, A. Tamkin, K. Nguyen, B. McLean, J. E. Burke, T. Hume, S. Carter, T. Henighan, and C. Olah. Towards monosemanticity: Decomposing language models with dictionary learning. https://transformer-circuits.pub/2023/monosemantic-features/index.html, 2023.

[29] H. Cunningham, A. Ewart, L. Riggs, R. Huben, and L. Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023.

[30] 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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 13782–13790, 2026. doi:10.1609/aaai.v40i16.38386.

[31] Z. Zhou, T. Cai, S. Z. Zhao, Y. Zhang, Z. Huang, B. Zhou, and J. Ma. AutoVLA: A vision-language-action model for end-to-end autonomous driving with adaptive reasoning and reinforcement fine-tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[32] D. Zhang, Z. Yuan, Z. Chen, C.-T. Liao, Y. Chen, F. Shen, Q. Zhou, and T.-S. Chua. Reasoning-VLA: A fast and general vision-language-action reasoning model for autonomous driving, 2025.

[33] NVIDIA Research. Alpamayo-R1: Bridging reasoning and action prediction for generalizable autonomous driving in the long tail, 2025.

[34] NVIDIA. NVIDIA Alpamayo: Open reasoning vision-language-action models for autonomous vehicle development. https://developer.nvidia.com/drive/alpamayo, 2026.

[35] R. C. Fong and A. Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3429–3437, 2017.

[36] R. Goldshmidt. Understanding what vision-language models see in traffic: Pixelshap for object-level attribution in autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1794–1802. IEEE Computer Society, 2025. doi:10.1109/ICCVW69036.2025.00189.

[37] M. Golovanevsky, W. Rudman, V. Palit, R. Singh, and C. Eickhoff. What do vlms notice? a mechanistic interpretability pipeline for gaussian-noise-free text-image corruption and evaluation, 2025. URL https://arxiv.org/abs/2406.16320.

[38] M. Pach, S. Karthik, Q. Bouniot, S. Belongie, and Z. Akata. Sparse autoencoders learn monosemantic features in vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[39] G. Jocher, A. Chaurasia, and J. Qiu. YOLOv8 by ultralytics. https://github.com/ultralytics/ultralytics, 2023.

[40] Google DeepMind. Gemini image generation and editing. https://deepmind.google/technologies/gemini/, 2024.

[41] B. F. Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

[42] V. Palit, R. Pandey, A. Arora, and P. P. Liang. Towards vision-language mechanistic interpretability: A causal tracing tool for blip. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 2856–2861, October 2023.

[43] B. Häon, K. Stocking, I. Chuang, and C. Tomlin. Mechanistic interpretability for steering vision-language-action models, 2025. URL https://arxiv.org/abs/2509.00328.

[44] C. Neo, L. Ong, P. Torr, M. Geva, D. Krueger, and F. Barez. Towards interpreting visual information processing in vision-language models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu, editors, *International Conference on Learning Representations*, volume 2025, pages 57172–57189, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/900fb3439e4968df79a6f2bfedec49cd-Paper-Conference.pdf.

[45] nostalgebraist. Interpreting gpt: The logit lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens, 2020. LessWrong. Accessed: 2026-05-29.

A Supplementary Tables and Figures

This appendix contains supporting visualisations and tables for the white-box and black-box analyses described in Sections 5.1 and 5.2.

A.1 Class-wise AD/FD Statistics

Table 3: Per-class object deviation statistics. Classes with $N < 15$ are aggregated into *Miscellaneous* (*tennis racket, motorcycle, sports ball, backpack, object, stop sign, handbag, parking meter, suitcase, bench, potted plant, chair*). Q1 and Q3 denote the 25th and 75th percentiles of AD.

Class	N	Mean AD	Q1 AD	Q3 AD	Max AD	Mean FD
Bus	47	0.663	0.085	0.763	3.931	1.889
Fire hydrant	17	0.540	0.117	1.003	2.037	1.454
Truck	183	0.505	0.089	0.703	3.548	1.419
Traffic light	531	0.482	0.071	0.624	4.534	1.399
Person	604	0.458	0.059	0.643	5.276	1.318
Bicycle	28	0.399	0.059	0.565	1.888	1.022
Car	1566	0.348	0.054	0.393	7.782	0.984
Umbrella	20	0.284	0.029	0.316	1.896	0.759
Misc. ($N < 15$)	66	0.403	0.130	0.580	1.588	1.154

A.2 Black-box Scene Visualizer

Figure 2: Scene Explorer — the interactive black-box visualisation dashboard for Counter_nuScenes. The left panel lists all objects in the scene ranked by AD, with gold/silver/bronze chips for the top-3 and colour-coded AD/FD bars; arrows indicate rank shift relative to the previous run. The right panel overlays per-object segmentation masks with bounding boxes and AD labels. The bottom bar shows scene-level statistics (object count, max AD, mean AD, top object). Multiple runs (Run-1, Run-2, Run-3, Merged) can be selected to compare rank orderings across seeds. Scene-0043 is shown: a pedestrian (AD 0.944 m) ranks first despite a large car occupying most of the frame, illustrating cases where physical scale does not predict causal influence.

A.3 Hidden-State Delta Curves

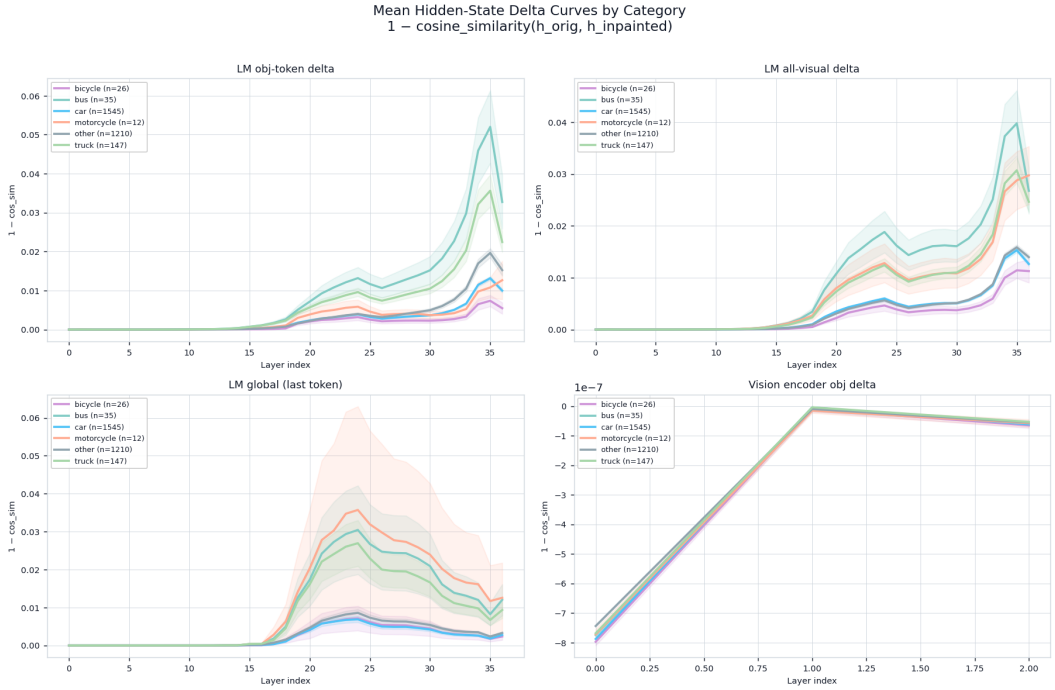

Figure 3: Mean hidden-state delta curves ($1 - \cos_{\text{sim}}$, original vs. inpainted) by object category across all LM layers and vision encoder blocks. **Top left**: object-token delta, averaged over patch positions spatially covering the removed object. **Top right**: all-visual delta, averaged over all 180 visual tokens; when this substantially exceeds the object-token delta, the removal is rippling into surrounding context. **Bottom left**: global delta at the last token (scene summary before the expert). **Bottom right**: vision encoder object delta at three probed blocks (8, 20, 26), scaled to 10^{-7}; the negative values indicate the inpainted region is more self-consistent than the original. Shaded bands show ± 1 standard error.

A.4 Per-Scene Z-Score Distribution

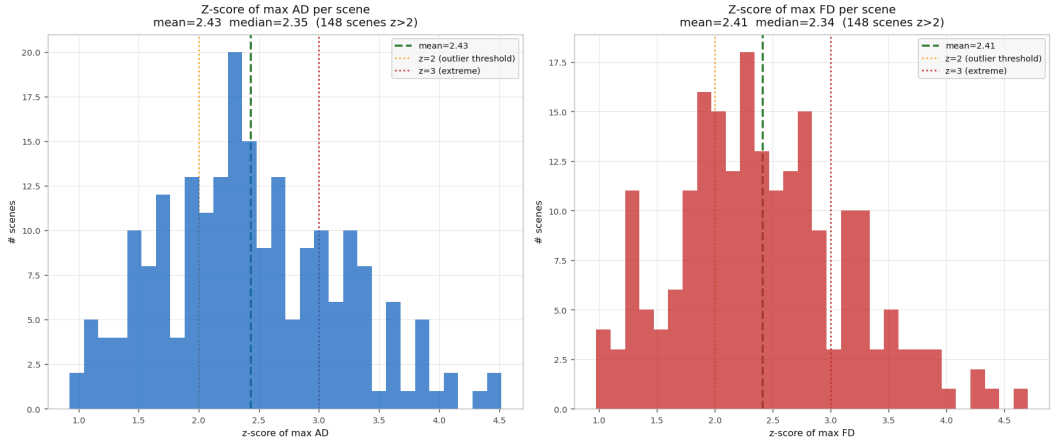

Figure 4: Distribution of the z-score of the maximum-AD object per scene, $z = (\max(\text{AD}) - \overline{\text{AD}})/\sigma_{\text{AD}}$, for both AD (left) and FD (right). Dashed lines mark $z = 2$ (strong outlier) and $z = 3$ (extreme outlier). 148 of 210 scenes (70.5%) exceed $z > 2$ for both metrics, indicating widespread single-object dominance across the dataset. See Section 5.1 for discussion.