Inside the Reasoning Menu: Discrete Failure Bands in Driving VLAs Under Sensor Perturbations

Abhinaw Priyadershi
NVIDIA Corporation
Santa Clara, USA
apriyadershi@nvidia.com

Jelena Frtunikj
NVIDIA GmbH
Munich, Germany
jfrtunikj@nvidia.com

***Abstract*—Vision-Language-Action (VLA) planners produce natural-language explanations alongside driving trajectories. Prior work found that when sensor perturbations change the explanation, trajectory error spikes by $5.3\times$, but the *magnitude* of text change has near-zero predictive power. We investigate the source of this mismatch. Using Alpamayo R1, a 10B-parameter driving VLA, we analyze the L2 trajectory errors of 5,443 changed-explanation pairs drawn from 15,968 evaluation pairs spanning 8 perturbation types. A Gaussian Mixture Model identifies six discrete severity bands, with BIC selecting k=6 (a sharp improvement from k=5 to k=6) and the preference remaining stable across 20 restarts. An action-transition analysis reveals that 71% of explanation changes are surface rewrites within the same action category; action labels do not significantly predict which severity band a failure lands in ($p = 0.306$, n=59 genuine switches). The structure is consistent with sub-categorical organization: geometric context (distance, timing, object proximity) determines the band, not the action word. For VLA pipelines that use Chain-of-Causation (CoC) text for downstream monitoring or fallback, binary CoC-change detection is therefore a more useful trigger than Jaccard-style magnitude scoring.**

***Index Terms*—VLA robustness, autonomous driving, discrete reasoning, trajectory multimodality, sensor perturbation**

I. Introduction

Vision-Language-Action (VLA) models bring a distinctive capability to autonomous driving: they produce natural-language explanations alongside predicted trajectories. Alpamayo R1 [1], a 10B-parameter open-weight driving VLA, outputs both a planned trajectory and a Chain-of-Causation (CoC) explanation, such as *"Slow down because the lead vehicle is braking ahead."* The VLA paradigm extends beyond driving: RT-2 [2] showed that vision-language-action models transfer web knowledge to robotic manipulation. In the driving domain, however, the CoC explanation is increasingly treated as a first-class signal: safety-oriented evaluation frameworks can inspect it for runtime auditing, and downstream modules can condition on it.

What happens to that signal under sensor perturbation? Priyadershi and Frtunikj [3] found that when perturbations change the CoC, trajectory error increases $5.3\times$ ($r_{\text{pb}} = 0.53$, Cohen's $d = 1.12$, $n = 15{,}968$). However, the *magnitude* of text change has near-zero predictive power ($r = -0.027$): a complete CoC rewrite and a single-word substitution produce statistically identical trajectory damage. That result shows that binary CoC change is informative, but continuous similarity is not. It does not, however, explain why this mismatch arises or what structure underlies the resulting trajectory errors. We address that question here.

The mismatch also exposes a gap in current robustness benchmarks [4], [5], which report mean trajectory error and attack success rates across perturbation conditions but do not decompose failures by severity. If trajectory outcomes are discrete, clustered into a small number of severity bands, such curves mask the actual failure structure. Recent NLP work has shown that LLMs form discrete concept clusters rather than smooth manifolds [6], and Chain-of-Thought prompting [7] enables structured step-by-step reasoning. Whether VLA trajectory outcomes exhibit the same discrete property is an open question.

We study whether changed-CoC cases are better described by movement among discrete severity bands than by smooth degradation. Concretely, we analyze 15,968 evaluation pairs from Alpamayo R1 under eight perturbation types and focus on the 5,443 pairs in which the explanation changes. We then test whether the resulting L2 trajectory errors exhibit clustered structure and whether that structure aligns with action-category changes.

Our contributions are threefold:

- We find evidence that changed-CoC trajectory errors are organized into six discrete severity bands, with GMM model selection favoring BIC k=6 and the preference remaining stable across 20 restarts.
- We show that many explanation changes are surface rewrites within the same action category, and our results are consistent with sub-categorical structure below coarse action labels.
- We show that, in this VLA evaluation setting, cluster-level outcome distributions provide a more informative robustness summary than mean degradation curves alone.

TABLE I
SCENARIO DISTRIBUTION AND CLEAN-BASELINE PERFORMANCE. FOLLOW_VEHICLE HAS THE HIGHEST COC FLIP RATE; TURN_RIGHT (n=40) SHOULD BE INTERPRETED WITH SMALL-SAMPLE CAUTION.

Scenario	n clips	Clean ADE (m)
FOLLOW_VEHICLE	475	1.75
INTERSECTION	344	2.12
STOP_SIGNAL	302	2.47
OTHER	418	2.26
PASSING	177	1.54
LANE_KEEPING	213	1.31
TURN_RIGHT	40	3.11

II. EXPERIMENTAL SETUP AND CORE RESULT

A. Model and Dataset

We evaluate Alpamayo R1 [1], a 10B-parameter driving VLA with open weights. At each inference step the model produces a 64-waypoint trajectory and a COC explanation in natural language (e.g., "Slow down because the lead vehicle is braking ahead"). The evaluation corpus consists of 1,996 real-world driving clips from the NVIDIA PhysicalAI-AV dataset [8], spanning seven scenario categories (Table I).

B. Perturbations and Metrics

We evaluate eight sensor perturbations spanning three corruption families: additive Gaussian noise ($\sigma \in \{10, 30, 50, 70\}$), photometric scaling (darkening $0.4\times$, brightening $1.6\times$), and volumetric scattering (fog at $\alpha \in \{0.3, 0.7\}$). Each family affects a different stage of the vision pipeline (sensor readout, exposure, atmospheric transmission). These perturbations are applied synchronously across all camera views. All perturbations operate at the pixel level; ego-pose and sensor metadata are unchanged. Each clip is tested under all eight conditions, yielding $1{,}996 \times 8 = 15{,}968$ (clip, attack) pairs.

For each pair, we record the trajectory displacement, a binary $\mathrm{coc\_changed}$ flag, and Jaccard token similarity $J \in [0, 1]$ between the clean and perturbed COC. Trajectory displacement is measured as the L2 distance between corresponding clean and perturbed trajectory waypoints, aggregated over all waypoints in the 64-step prediction horizon. Section III tests whether the L2 distribution of changed-COC pairs is discrete or continuous.

C. The Binary Signal and the Dimmer Paradox

Of the 15,968 pairs, 10,525 (66%) preserve the COC text and have mean L2 ≈ 4.1 m. The remaining 5,443 (34%) change the COC (mean L2 $= 21.8$ m), corresponding to $5.3\times$ increase in trajectory error. The point-biserial correlation between the binary $\mathrm{coc\_changed}$ flag and L2 displacement is $r_{\mathrm{pb}} = 0.53$ (Cohen's $d = 1.12$, $p \approx 0$). At the perturbation level, the COC change rate and mean L2 deviation correlate at $r = 0.994$ across all eight conditions [3].

In contrast, the magnitude of the text change carries little information about error severity. Among the 5,443 changed pairs, Jaccard token similarity and L2 displacement correlate at $r = -0.027$ (Spearman $\rho = 0.010$): a complete COC rewrite ($J < 0.10$, n=556) and a mid-range edit ($0.10 \le J \le 0.90$, n=4,731) produce statistically indistinguishable trajectory damage (Mann-Whitney $p = 0.257$, Cohen's $d = -0.026$; Fig. 1). Post-hoc power exceeds 99% to detect $d \ge 0.10$, ruling out any continuous gradient larger than $d \approx 0.06$.

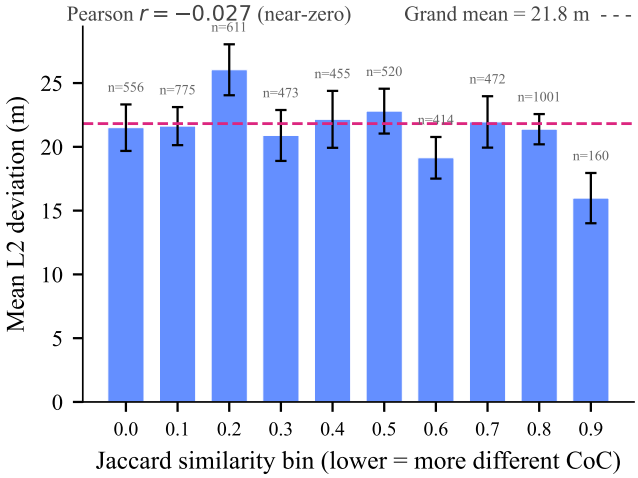

Fig. 1. The dimmer paradox. Mean L2 trajectory error per Jaccard similarity bin (changed pairs, n=5,443). Every bin hovers near the grand mean of 21.8 m regardless of how much the text changed ($r = -0.027$). Complete rewrites ($J < 0.1$) and minor edits ($J > 0.8$) produce equivalent damage.

We refer to this as the *dimmer paradox*: the binary COC-change flag predicts trajectory damage, but the magnitude of the change does not. The natural-language explanation tells you *that* the model switched, but not *how far* it deviated. This result motivates the next section, which tests whether changed-COC trajectory outcomes are organized into discrete severity bands.

For VLA pipelines that use COC text for downstream safety monitoring, the near-zero Jaccard/L2 correlation suggests that magnitude-based metrics (including embedding distance) are unlikely to outperform binary change detection for trajectory-damage prediction in this setting. A binary COC-change detector is both simpler and more effective on this data.

III. THE DISCRETE REASONING MENU

Section II showed that binary COC change is informative, whereas the magnitude of the text change is not. We now test whether trajectory outcomes for changed-COC pairs are better described by a small number of discrete severity bands than by smooth degradation.

A. GMM Analysis

A Gaussian Mixture Model [9] fitted to the L2 distribution of the 5,443 changed-COC pairs yields BIC-optimal k=6 components, selected in all 20 random restarts. The BIC improvement from k=5 to k=6 is larger than the improvement from adding a 7th or 8th component, and five-fold cross-validation over k=1, ..., 8 likewise favors k=6 in all folds. Table II summarizes the six components, and Fig. 2 (center) visualizes them.

As a concrete illustration: noise_70 and bright produce similar mean L2 errors among changed pairs (23.3 m vs. 21.0 m), but noise_70 places 7.5% of failures in the catastrophic C4/C5 bands versus 4.9% for bright. Reporting only the mean hides a 53% difference in catastrophic-event exposure.

TABLE II
GMM k=6 CLUSTER STRUCTURE ON CHANGED PAIRS (n=5,443). COMPONENT MEANS ARE FIXED GLOBALLY FROM THE FULL CORPUS.

Band	Mean L2	Share	Interpretation
C0	5.3 m	35%	Near-stop; minor wobble
C1	16.0 m	31%	Creep; noticeable drift
C2	29.4 m	17%	Follow-adjustment
C3	47.2 m	11%	Moderate maneuver
C4	76.0 m	4%	Hard maneuver
C5	124.2 m	1%	Emergency

TABLE III
CONVERGENT EVIDENCE BATTERY (5,443 CHANGED PAIRS). MOST ANALYSES ARE CONSISTENT WITH THE DISCRETE-BAND INTERPRETATION. COC LENGTH IS A NEUTRAL RESULT: IT PREDICTS FRAGILITY BUT DOES NOT DISTINGUISH BETWEEN DISCRETE AND CONTINUOUS DEGRADATION MODES.

Analysis	Result	Discrete?
Bimodality	BC $= 0.642 > 0.555$ [11]	Yes
Jaccard/L2	$r = -0.027$ (near-zero)	Yes
Cross-attack	$\rho = 0.347$ (clip-level)	Yes
CoC length	$r = 0.246$, $p < 0.001$	Neutral
Noisy proxy	$d = -0.026$, $p = 0.257$	Yes
Damage equiv.	$d = 0.076$; low=high Jaccard	Yes
Clip-level ICC	ICC $= 0.35$ (between-clip)	Yes
GMM BIC	k=6, sharp gap at k=5→6	Yes

Per-scenario GMMs (k=1 . . . 6) select k=2 to k=5 within each scenario; no individual scenario is unimodal. Cramér's $V = 0.135$ ($\chi^2(45) = 498.7$, $p < 0.001$) confirms that cluster assignment is largely scenario-independent: the global k=6 does not merely reflect six scenario types each contributing one characteristic error value. We treat the k=6 structure as an observable property of the L2 distribution, not as a claim about internal VLA decision mechanisms.

B. Convergent Evidence

Table III summarizes several complementary analyses that are broadly consistent with the discrete-band interpretation. The most diagnostic additional test is the *noisy proxy*: mid-range Jaccard values (0.10 to 0.90, accounting for 86.9% of changed pairs) produce statistically identical damage as complete rewrites ($J < 0.10$): Cohen's $d = -0.026$, $p = 0.257$. These mid-range texts are not intermediate-severity failures. They are paraphrases of the same discrete decision expressed in different words.

Cross-attack stability ($\rho = 0.347$, BC $= 0.607$) indicates that fragility is a clip-level property (ICC $= 0.35$ [10]): clips vulnerable to one attack type tend to be vulnerable to others. A full 26.5% of clips never flip under any of the eight attacks; 2.5% always flip.

C. Surface Rewrites and Sub-Categorical Structure

This subsection draws on two scales of evidence: the full 5,443-pair corpus via rule-based action extraction (94.8% agreement with manual labels on non-OTHER categories), and a 200-pair manually annotated subset for the action-category equivalence test.

The transition matrix (Fig. 2, right), computed on all 5,443 pairs, reveals that 71% of CoC changes preserve the original action category (per-action range: 56% for KEEP-LANE to 78% for STOP/FOLLOW/PASS). The model rewrites its explanation in different words, but the action category typically remains unchanged.

Do action labels predict which severity band a failure lands in? On the 200-pair manually annotated subset, we labeled primary driving actions (STOP, FOLLOW, PASS, YIELD, KEEP-LANE, TURN). Action-transition pairs produce significantly different L2 levels (Kruskal-Wallis $p < 0.0001$): this indicates that coarse action labels capture real variation in trajectory outcomes. However, surface rewrites (same action, different words; n=141, mean L2 $= 60.6$ m) and genuine category switches (n=59, mean L2 $= 51.2$ m) are not significantly different (Fig. 3; MWU $p = 0.306$, $d = 0.20$; TOST equivalence [12] at $\Delta = 24.1$ m: $p = 0.019$).

Taken together, these results are consistent with a *sub-categorical* structure below coarse action labels. Pairs sharing the same action word can still fall into different severity bands, suggesting that factors such as geometric context, timing, and object proximity may matter more than action category alone. A tabular Random Forest trained on clip-level metadata (scenario type, perturbation type, severity level) achieves macro-F1 $= 0.354$ (2.1× above chance) for cluster prediction but RMSE $= 16.1$ m on all pairs, worse than the binary baseline B1 (13.6 m). This suggests that the observed structure is not easily recovered from the available metadata features alone.

Summary. Four lines of evidence converge: GMM BIC k=6 (sharp gap, 20 restarts); noisy proxy ($d = -0.026$); cross-attack stability ($\rho = 0.347$); sub-categorical structure ($p = 0.306$). In this evaluation, the VLA's trajectory-outcome space is better described as a discrete menu than a continuous manifold.

IV. DISCUSSION AND FUTURE WORK

A. Implications for VLA Pipeline Evaluation

The discrete six-band structure has a practical implication for the robustness evaluation: two perturbation types with identical mean trajectory error can carry different risk. One may concentrate failures in C0/C1 (safe), while the other places a disproportionate share in the catastrophic C4/C5 bands. As a result, mean degradation curves alone may obscure safety-relevant differences between perturbation conditions.

The near-zero Jaccard/L2 relationship ($r = -0.027$) suggests that magnitude-based text similarity metrics are not a reliable proxy for trajectory-damage severity for this model in this evaluation setting. In contrast, binary CoC-change detection is useful not because it is inherently superior, but because the underlying phenomenon in this evaluation is better described as a discrete switch than as a continuous gradient.

We therefore propose a cluster-level evaluation step for VLA pipelines. Given a new perturbation condition: (1) run the VLA on n perturbed clips, (2) fit a GMM to the L2 distribution of changed-CoC pairs, (3) report the C4+C5 proportion as the catastrophic-event rate. Under this view, two perturbation conditions with similar mean error but different C4/C5 rates should receive different safety assessments (as in the noise_70

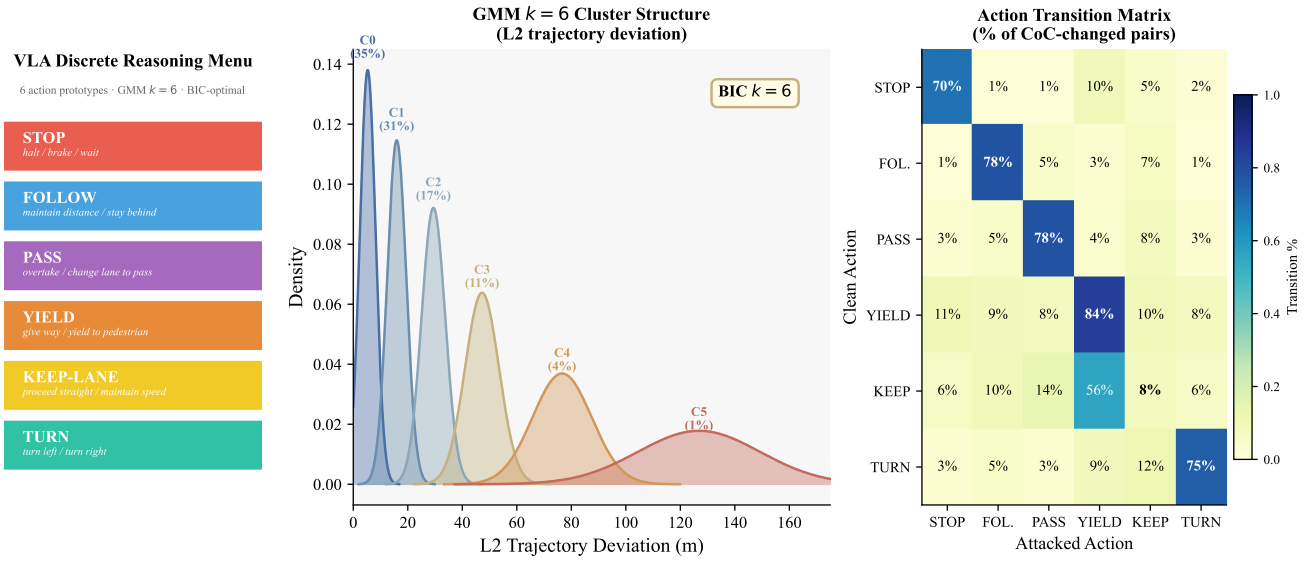

Fig. 2. The Discrete Reasoning Menu. **Left:** six action prototypes extracted from Alpamayo R1 CoC explanations (GMM k=6, BIC-optimal). **Center:** GMM cluster structure on L2 trajectory deviations of changed pairs (n=5,443); the sharp BIC kink at k=5 → 6 confirms genuine discrete structure. **Right:** action transition matrix (real data, n=5,310 non-OTHER pairs) showing that 56 to 78% of CoC changes per action category are surface rewrites (diagonal dominance).

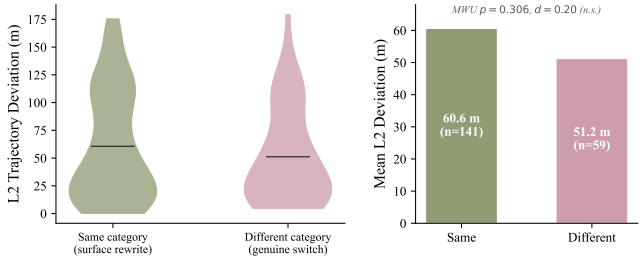

Fig. 3. Sub-categorical equivalence on the manually annotated subset (n=200). **Left:** L2 distributions for surface rewrites (same action category, n=141) and genuine switches (different category, n=59) overlap substantially. **Right:** mean L2 is not significantly different (MWU $p = 0.306$, $d = 0.20$). These results are consistent with, but do not by themselves establish, sub-categorical structure.

vs. bright comparison in Section III). This analysis is post hoc and requires no additional model training.

B. Limitations

This study evaluates a single VLA (Alpamayo R1) on a single dataset of 1,996 clips under eight pixel-level perturbations. Temporal corruptions (e.g. frame drops, sensor latency) and adversarial patches are untested.

The sub-categorical analysis is also limited by annotation scale. The manually annotated subset contains 200 pairs, including 59 action-category switches, which limits the statistical power for detecting small effects. Accordingly, the evidence supports a cautious interpretation: the observed structure is consistent with sub-categorical organization, but does not yet establish its generality (i.e. for other VLA architectures and datasets).

More broadly, the six-component GMM should be interpreted as a useful summary of the observed L2 distribution in this setting, not as a direct measurement of latent internal decision states. A second open-weight VLA with a unimodal error distribution under the same protocol would weaken the case for the present interpretation.

C. Future Work

Alpamayo R1.5, recently released with updated training and architecture, offers an immediate opportunity for same-family cross-version testing. Senna [13] and other open-weight driving VLAs are candidates for cross-family replication. A second direction is to broaden the perturbation beyond pixel-level corruption. Extending the perturbation to temporal corruptions and richer weather simulation beyond pixel-level fog would test whether the observed six-band structure is a general property of VLA failure or specific to the current evaluation setting. Finally, the observed discrete structure suggests a path toward severity-graded runtime monitoring, in which cluster posterior probabilities could support a multi-tier response policy; we leave this to future work.

V. Conclusion

In this evaluation setting, VLA trajectory outcomes under sensor perturbation are better described by discrete severity bands than by a smooth degradation model. Across 15,968 evaluation pairs, changed-CoC cases exhibit a strong binary association with trajectory error, while the magnitude of the text change carries near-zero predictive signal.

A GMM analysis of the 5,443 changed-CoC pairs identifies a six-band error structure that is stable across restarts and supported by additional convergent analyses. The action-transition results are consistent with a sub-categorical interpretation, although broader replication is still required.

The practical implication is twofold. First, binary CoC-change detection is a stronger and simpler pipeline trigger than magnitude-based scoring for trajectory-damage prediction. Second, mean degradation curves can mask safety-critical differences between perturbation conditions; cluster-level outcome distributions (particularly the C4+C5 catastrophic-event rate) provide a more informative evaluation signal.

REFERENCES

[1] Y. Wang, W. Luo, J. Bai, Y. Cao, T. Che, K. Chen, Y. Chen *et al.*, "Alpamayo-R1: Bridging reasoning and action prediction for generalizable autonomous driving in the long tail," arXiv:2511.00088, 2025.

[2] A. Brohan *et al.*, "RT-2: Vision-language-action models transfer web knowledge to robotic control," in *Proceedings of the Conference on Robot Learning (CoRL)*, 2023, arXiv:2307.15818.

[3] A. Priyadershi and J. Frtunikj, "Lost in fog: Sensor perturbations expose reasoning fragility in driving VLAs," in *Proc. Workshop on Safe AI for Autonomous Driving (SAIAD), CVPR*, 2026, accepted; to appear.

[4] L. Kong *et al.*, "The RoboDrive challenge: Drive anytime anywhere in any condition," arXiv:2405.08816, 2024.

[5] T. Zhang, L. Wang, X. Zhang *et al.*, "Visual adversarial attack on vision-language models for autonomous driving," arXiv:2411.18275, 2024.

[6] J. Mamou, H. Le, M. Del Rio, C. Stephenson, H. Tang, Y. Kim, and S. Chung, "Emergence of separable manifolds in deep language representations," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2020, arXiv:2006.01095.

[7] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[8] NVIDIA, "PhysicalAI-Autonomous-Vehicles evaluation dataset," https://huggingface.co/datasets/nvidia/PhysicalAI-Autonomous-Vehicles, 2025.

[9] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[10] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *Journal of Chiropractic Medicine*, vol. 15, no. 2, pp. 155–163, 2016.

[11] R. Pfister, K. A. Schwarz, M. Janczyk, R. Dale, and J. B. Freeman, "Good things peak in pairs: A note on the bimodality coefficient," *Frontiers in Psychology*, vol. 4, p. 700, 2013.

[12] D. Lakens, "Equivalence tests: A practical primer for t tests, correlations, and meta-analyses," *Social Psychological and Personality Science*, vol. 8, no. 4, pp. 355–362, 2017.

[13] B. Jiang, S. Chen, B. Liao, X. Zhang, W. Yin, Q. Zhang, C. Huang, W. Liu, and X. Wang, "Senna: Bridging large vision-language models and end-to-end autonomous driving," arXiv:2410.22313, 2024.