Task-Aligned Stability Analysis of Vision-Language Models for Autonomous Driving Hazard Detection

Everett Richards [1]

Abstract

Vision-language models (VLMs) are increasingly used for scene understanding in autonomous driving, but robustness analysis often relies on task-agnostic embedding stability alone. We study whether corruption-induced embedding drift predicts changes in a task-aligned hazard score derived from CLIP image-text similarities. Using controlled corruptions on BDD100K road scenes, we compare embedding drift against margin drift, defined as the change in hazard score under perturbation. The relationship is highly corruption-dependent: some families exhibit strong coupling between representation drift and decision drift, while others induce hazardous decision instability despite relatively modest embedding change. Furthermore, corruption families differ in failure direction: most suppress hazard detections via false negatives, while occlusion instead triggers false alarms, suggesting that benchmark design should account for asymmetric failure modes, not just overall instability rates. These results suggest that robustness benchmarks should include task-aligned stability measures in addition to embedding-level perturbation statistics.

1. Introduction

Autonomous Vehicles (AVs) are becoming increasingly popular, and are already being deployed in many cities around the world (Litman, 2025). However, public confidence in self-driving technology remains astonishingly low, with a 2025 survey suggesting that only 38% of Americans would be willing to ride in a self-driving car if given the opportunity (Wali, 2026). Furthermore, a 2024 study demonstrated that people are more inclined to blame a self-driving car than a human-operated car for a collision, even in otherwise equivalent scenarios (Zhang et al., 2024).

[1]Department of Computer Science, San Diego State University. Correspondence to: Everett Richards <ehrichards9@gmail.com>.

Accepted to the 1st Workshop on Combining Theory and Benchmarks, CTB@ICML 2026, Seoul, South Korea. Copyright 2026 by the author(s).

In recent years, researchers have identified perception as an essential safety factor in autonomous driving (Richards et al., 2025b). Autonomous vehicles depend on perception systems that remain reliable under distribution shift. Hazard detection is a particularly safety-critical instance of this problem because small visual changes can alter the driving decision implied by a scene. Vision-language models (VLMs) are attractive for this setting because they combine visual recognition with prompt-based semantic reasoning, enabling natural-language descriptions of hazardous and non-hazardous traffic scenarios (Bordes et al., 2024; Xia et al., 2026). However, prior work shows that VLM-based hazard detectors can be brittle, overreact in ambiguous settings, and impose substantial computational cost (Choi et al., 2025; Adil et al., 2025).

A common robustness perspective is to measure how much the image embedding changes under corruption. While useful, this quantity is task-agnostic: a large change in embedding space need not produce a large change in the decision variable that matters for safety. We therefore contrast *embedding drift* with a task-aligned *margin drift* based on a CLIP hazard score. This aligns with broader efforts toward quantitative reliability analysis, robustness under distribution shift, and principled benchmark design.

Whereas prior works have explored the effects of data perturbations on multi-modal object detection pipelines (Richards et al., 2025a) and reinforcement learning-based action trajectory planning (Richards & Dai, 2025), this paper aims to model the degradation of VLM-based hazard detection and contribute to our understanding of notable failure cases.

Our central question is simple: *to what extent does corruption-induced embedding drift predict decision-relevant instability in VLM hazard detection?* We find that the answer depends strongly on corruption family. JPEG compression and downsampling produce relatively predictable behavior, whereas motion blur and occlusion can trigger high decision flip rates without proportionally large embedding movement. This suggests that representation stability alone is an incomplete proxy for safety-relevant robustness.

2. Method

Hazard scoring. Let $\hat{f}(x)$ and $\hat{g}(t)$ denote frozen CLIP image and text encoders (both normalized). We define hazardous prompts $\mathcal{H}$ and non-hazardous prompts $\mathcal{N}$, then score image x by

$$\Lambda(x) = \max_{h \in \mathcal{H}} \langle \hat{f}(x), \hat{g}(h) \rangle - \max_{n \in \mathcal{N}} \langle \hat{f}(x), \hat{g}(n) \rangle, \quad (1)$$

where larger values indicate greater hazard likelihood. Positive scores correspond to hazardous scenes and negative scores to non-hazardous scenes.

Our prompt sets are intentionally simple and interpretable. Hazardous prompts include phrases such as *a pedestrian crossing the road*, *a cyclist in the lane*, and *a stopped vehicle blocking traffic*. Non-hazardous prompts include *a clear road with no obstacles*, *a normal highway scene*, and *a clear lane ahead*. This follows the prompt-based hazard scoring strategy used in recent autonomous-driving work built on CLIP (Radford et al., 2021; Greer et al., 2026).

Corruptions and drift metrics. For corruption family c and severity $\alpha \in \{1, \ldots, 5\}$, let $T_{c,\alpha}(x)$ denote the corrupted image. We evaluate seven families: fog, Gaussian blur, motion blur, JPEG compression, low light, occlusion, and downsampling. We then define

$$\Delta(x, c, \alpha) = \|\hat{f}(x) - \hat{f}(T_{c,\alpha}(x))\|_2 \quad (2)$$

and

$$D_\Lambda(x, c, \alpha) = |\Lambda(x) - \Lambda(T_{c,\alpha}(x))|. \quad (3)$$

Here Δ is task-agnostic embedding drift and D_Λ is task-aligned margin drift.

We also record a binary *flip* indicator,

$$\texttt{flip} = \mathbb{1}[\text{sign}(\Lambda(x)) \neq \text{sign}(\Lambda(T_{c,\alpha}(x)))], \quad (4)$$

which captures whether corruption changes the hazard verdict.

Cauchy-Schwarz upper bound. Suppose that the most-aligned prompts $h^* \in \mathcal{H}$ and $n^* \in \mathcal{N}$ are stable under perturbation, i.e. that their relative order is unchanged by corruption operators. Then the Λ formulation in Eq. 1 collapses to:

$$\Lambda(x) = \langle \hat{f}(x), \hat{g}(h^*) - \hat{g}(n^*) \rangle \quad (5)$$

Then, the Cauchy-Schwarz inequality gives us an upper bound on margin drift D_Λ, where the coefficient L depends on the prompt sets $\mathcal{H}, \mathcal{N}$:

$$|D_\Lambda| \leq \|\hat{g}(h^*) - \hat{g}(n^*)\|_2 \cdot \|\hat{f}(x) - \hat{f}(T(x))\|_2 = L \cdot \Delta \quad (6)$$

Algorithm 1 Task-aligned corruption evaluation

Require: Images $\mathcal{X}$, corruptions $\mathcal{C}$, severities $\mathcal{A}$, prompt collections $\mathcal{H}, \mathcal{N}$

```
for c ∈ C do
    for α ∈ A do
        for x ∈ X do
            Compute Λ(x) and f̂(x)
            Form x' = T_{c,α}(x)
            Compute Λ(x') and f̂(x')
            Set Δ = ‖f̂(x) − f̂(x')‖_2
            Set D_Λ = |Λ(x) − Λ(x')|
            Set flip = 1[sign(Λ(x)) ≠ sign(Λ(x'))]
        end for
    end for
end for
```

3. Experimental setup

We use 2,000 images sampled from the BDD100K validation split (Yu et al., 2020). For each image, corruption family, and severity, we compute the metrics above and then summarize each family using: (i) Spearman correlation $\rho(\Delta, D_\Lambda)$, (ii) flip rate, and (iii) mean drift as a function of severity. This design isolates the relationship between representation change and decision instability under controlled shift.

4. Results

4.1. Hazard Score Behavior

The CLIP-based score produces plausible rankings: low-scoring scenes are typically open roads with minimal interaction, while high-scoring scenes contain dense traffic, nearby actors, or explicit obstructions. This supports the use of Λ as a continuous decision variable for analyzing corruption effects, even in the absence of a fully supervised detection benchmark. See Appendix B for a visual illustration of Λ correctly approximating hazard presence in real-world driving scenes.

4.2. Embedding Drift and Margin Drift Are Correlated, but Not Interchangeable

Table 1 shows strong heterogeneity across corruption families. JPEG compression and downsampling have relatively strong rank correlation between embedding drift and margin drift, suggesting that representation movement is a reasonable proxy for decision movement in those cases. In contrast, occlusion and fog show much weaker coupling. Most notably, motion blur exhibits the highest flip rate even though its drift correlation is only moderate. Thus, the corruption families that most threaten decision reliability are not necessarily those that induce the largest embedding perturbations.

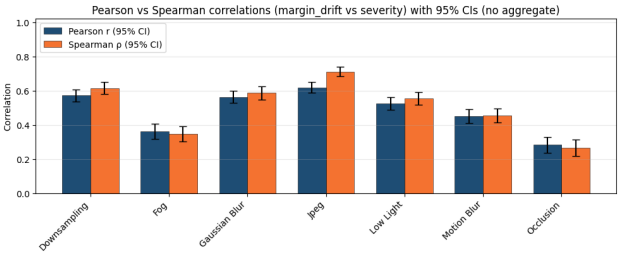


Figure 1. Family-wise comparison of Spearman correlation coefficient ρ and Pearson correlation coefficient r, relating embedding drift Δ to margin drift D_Λ. Spearman values close to 1 indicate a strong *monotonic* relationship, whereas Pearson values close to 1 suggest a strong *linear* relationship.

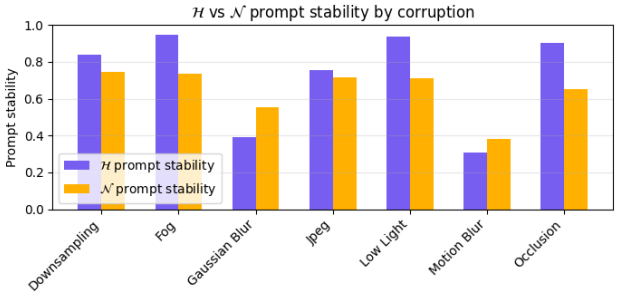
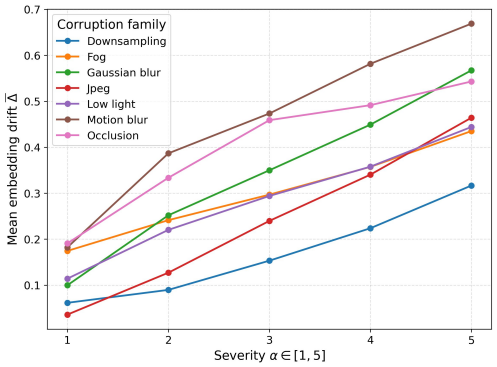


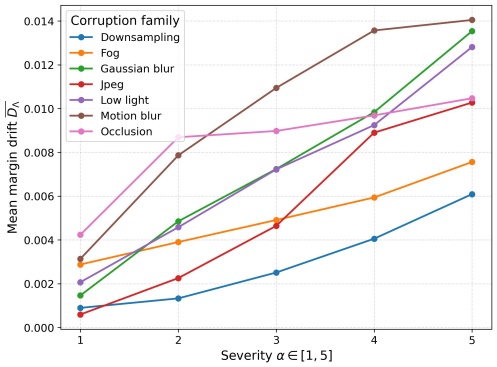
Figure 2. Family-wise comparison of $\mathcal{H}$ prompt stability and $\mathcal{N}$ prompt stability, for severity $\alpha = 5$. Values close to 1 indicate that the leading prompt is largely invariant to the corruption operator.

Fig. 1 illustrates the Spearman (rank) correlation coefficient ρ and Pearson (linear) correlation coefficient r for each corruption family. The two correlation metrics are tightly coupled and weakly to moderately strong across each corruption family, indicating an imperfect but non-negligible monotone and linear relationship between embedding drift and margin drift. Fig. 2 provides a cross-family comparison of prompt stability, which measures the proportion of images whose most closely-aligned prompt in each prompt set ($\mathcal{H}$ or $\mathcal{N}$) is unchanged upon corruption. Values close to 1 support the prompt order stability assumption required by Eq. 6, while smaller values violate this assumption. In either case, this theoretical upper bound cannot be guaranteed due to the nonlinearity in $D_\Lambda(x, c, \alpha)$ (with respect to α) induced by a change in leading prompt.

Figure 3 visually summarizes the roughly-monotone, roughly-linear relationships between Δ vs. α and D_Λ vs. α. The left panel plots mean embedding drift against severity, and the right panel plots mean margin drift against severity. Both generally increase monotonically, confirming that stronger corruptions perturb both representation and task score. However, the relative ordering differs across families. Motion blur and occlusion induce disproportionately large margin effects relative to their embedding movement, while JPEG compression causes substantial embedding change with comparatively lower decision instability.



(a) Mean embedding drift vs. severity.



(b) Mean margin drift vs. severity.

Figure 3. Severity trends differ across corruption families. Similar monotonicity does not imply similar safety impact.

4.3. Scatter Geometry of Embedding and Margin Drift

To further analyze the relationship between representation change and decision instability, we examine the joint distribution of embedding drift Δ and margin drift D_Λ across corruption families. Figure 4 visualizes this relationship. Several consistent structural patterns emerge:

Upper-envelope (conical) constraint. Across all corruption families (Fig. 4a), the data occupies a cone-shaped region with a clear upper boundary relating D_Λ to Δ. Large margin shifts are only observed when embedding drift is non-negligible, suggesting that representation change is a necessary condition for decision change. However, the spread within this cone is substantial, indicating that embedding drift alone does not determine the magnitude of decision impact.

Family-dependent coupling regimes. Different corruption families populate distinct regions of this space. JPEG compression (Fig. 4b) exhibits a tight, approximately linear relationship between Δ and D_Λ, consistent with its high Spearman correlation in Table 1. In contrast, motion blur and occlusion (Fig. 4c–d) display much broader vertical dispersion. For these families, similar levels of embedding drift

Table 1. Family-wise coupling between embedding drift and decision drift, with 95% confidence intervals on Spearman (ρ) and Pearson (r) correlation coefficients. FPR and FNR denote negative-to-positive and positive-to-negative hazard verdict flips, respectively. $\mathcal{H}$ and $\mathcal{N}$ stability measure the proportion of images whose leading prompt is unchanged under corruption.

Corruption family	$\rho(\Delta, D_\Lambda)$	$r(\Delta, D_\Lambda)$	Flip rate (%)	FPR (%)	FNR (%)	$\mathcal{H}$ stability (%)	$\mathcal{N}$ stability (%)
Downsampling	0.615 ± 0.035	0.572 ± 0.034	5.9	1.8	4.1	84.2	74.7
Fog	0.348 ± 0.044	0.360 ± 0.044	10.8	5.8	5.1	94.6	73.4
Gaussian blur	0.587 ± 0.038	0.563 ± 0.035	16.2	2.7	13.5	39.4	55.6
JPEG compression	0.711 ± 0.027	0.618 ± 0.031	10.8	0.7	10.2	75.8	71.7
Low light	0.553 ± 0.037	0.524 ± 0.037	15.4	3.4	12.0	93.6	71.0
Motion blur	0.455 ± 0.042	0.451 ± 0.041	24.2	4.2	20.1	30.6	38.0
Occlusion	0.266 ± 0.049	0.283 ± 0.047	17.8	15.3	2.5	90.2	65.3
Average	0.473 ± 0.015	0.438 ± 0.016	14.4	4.8	9.6	72.6	64.3

can correspond to widely varying margin shifts, indicating weaker coupling between representation and decision.

Low-drift, high-impact regime. Of particular concern is the region characterized by relatively small Δ but large D_Λ. This regime appears prominently in motion blur and occlusion. Points in this region correspond to perturbations that minimally affect the global embedding while significantly altering the relative similarity to hazardous versus non-hazardous prompts. Empirically, these cases account for a large fraction of observed flips, suggesting that they represent a critical failure mode for VLM-based hazard detection.

Severity stratification. Coloring by severity reveals that embedding drift increases relatively smoothly with corruption strength, while margin drift exhibits more heterogeneous growth. In tightly coupled families (e.g., JPEG), increasing severity moves points outward along a narrow band. In weakly coupled families (e.g., occlusion), severity increases both the mean and the variance of margin drift, further amplifying unpredictability at higher corruption levels.

Interpretation. The scatter structure indicates that embedding drift constrains but does not determine decision change. While large margin shifts require non-negligible representation movement, similar levels of embedding drift can lead to substantially different margin outcomes depending on corruption type. This asymmetry highlights the role of corruption-specific effects in shaping decision stability.

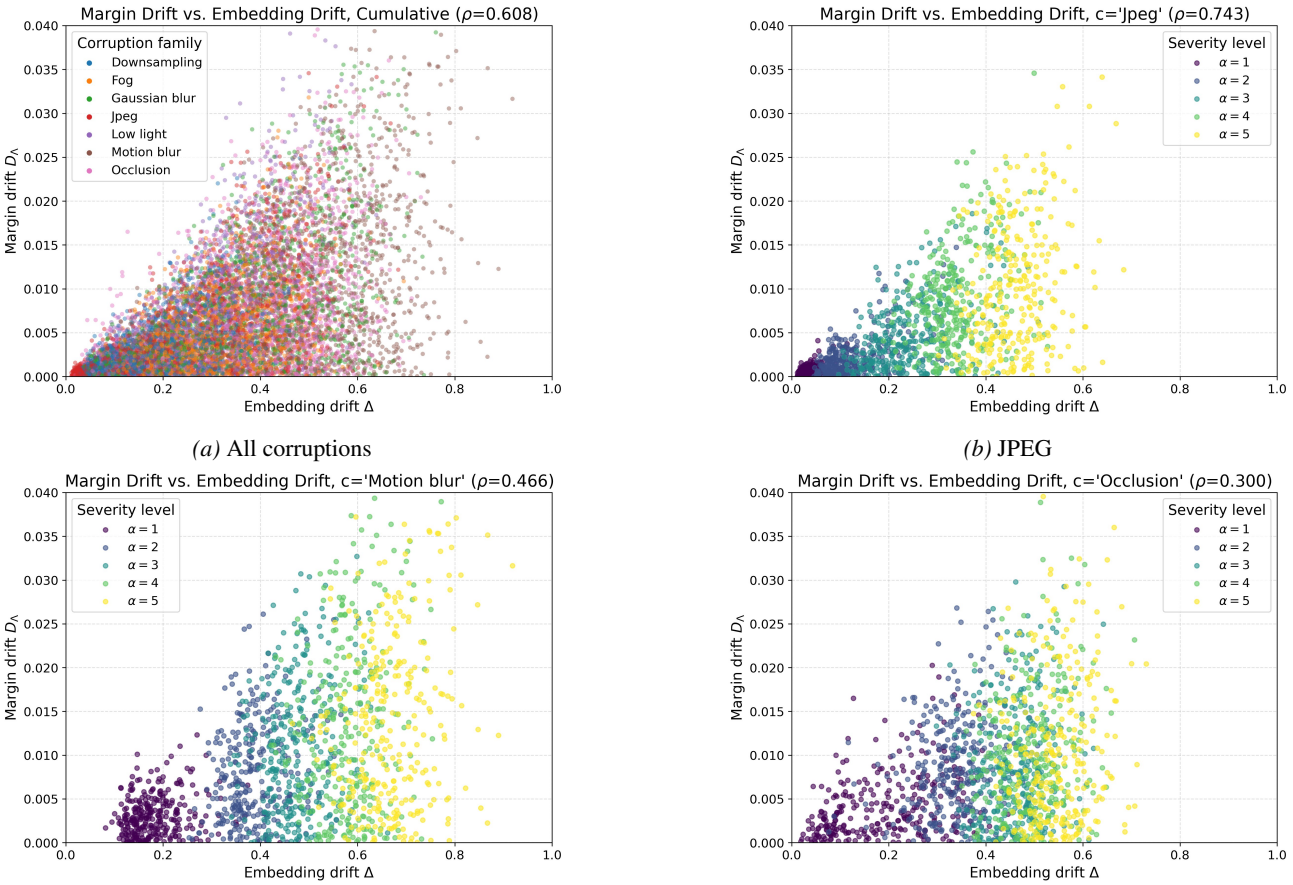
Hypothesis: bounded decision sensitivity. The observed upper-envelope structure suggests a potential Lipschitz-like relationship between embedding drift and margin drift of the form $D_\Lambda \leq L \cdot \Delta$ for some task-dependent constant L. Our empirical geometry indicates that decision changes are indeed constrained by representation movement but not tightly determined by it. Preliminary experiments using the Cauchy-Schwarz inequality (see Eq. 6) estimated the value of L to be somewhere in the range $[0.70, 1.10]$, depending on the choice of prompts, whereas visual inspection of Fig. 4 suggests that the boundary likely lies somewhere in the range of $[0.05, 0.07]$. This suggests that the true slope L of the tightest upper bound is an order of magnitude shallower than the Cauchy-Schwarz bound we had hypothesized. Future work could aim to determine this tight upper-envelope constant analytically or empirically, potentially yielding formal guarantees on decision stability under bounded embedding perturbations.

4.4. Implications for Benchmarking

These results suggest that robustness evaluation for multimodal systems should incorporate task-aligned metrics in addition to representation-level measures. In particular, reporting margin drift and threshold-flip rates by corruption family provides a more direct view of decision stability under shift. Benchmarks that rely solely on embedding perturbation statistics may overlook failure modes that have limited representation impact but significant decision consequences.

5. Conclusion

We presented a compact, task-aligned stability analysis for CLIP-based hazard detection under controlled image corruption. Across seven corruption families, embedding drift and hazard-margin drift were related but far from equivalent. Some families produced predictable coupling between the two, while others created dangerous decision instability that was not well reflected by embedding movement alone. Additionally, corruption families differ not only in flip rate but in failure direction: most families are dominated by false negatives that suppress hazard detections, while occlusion produces predominantly false positives — a divergence with direct implications for how corruptions should be weighted in safety-critical benchmarks. For theory-informed benchmark design, this suggests that representation robustness should be complemented by task-aligned stability metrics that quantify how corruption propagates into decision variables and final verdicts.



(a) All corruptions

(b) JPEG

(c) Motion blur

(d) Occlusion

Figure 4. Embedding drift vs. margin drift across corruption families. Points are colored by corruption family in (a) and by severity level in (b)–(d).

References

Adil, M., Lee, G., Gonzalez, V. A., and Mei, Q. Using vision language models for safety hazard identification in construction, 2025. URL https://arxiv.org/abs/2504.09083.

Bordes, F., Pang, R. Y., Ajay, A., Li, A. C., Bardes, A., Petryk, S., Mañas, O., Lin, Z., Mahmoud, A., Jayaraman, B., Ibrahim, M., Hall, M., Xiong, Y., Lebensold, J., Ross, C., Jayakumar, S., Guo, C., Bouchacourt, D., Al-Tahan, H., Padthe, K., Sharma, V., Xu, H., Tan, X. E., Richards, M., Lavoie, S., Astolfi, P., Hemmat, R. A., Chen, J., Tirumala, K., Assouel, R., Moayeri, M., Talattof, A., Chaudhuri, K., Liu, Z., Chen, X., Garrido, Q., Ullrich, K., Agrawal, A., Saenko, K., Celikyilmaz, A., and Chandra, V. An introduction to vision-language modeling, 2024. URL https://arxiv.org/abs/2405.17247.

Choi, D., Lee, S., and Song, Y. Better safe than sorry? overreaction problem of vision language models in visual emergency recognition, 2025. URL https://arxiv.org/abs/2505.15367.

Greer, R., Keskar, M., Martinez-Sanchez, A., Roy, P., Shriram, S., and Trivedi, M. Vision and language: Novel representations and artificial intelligence for driving scene safety assessment and autonomous vehicle planning, 2026. URL https://arxiv.org/abs/2602.07680.

Litman, T. Autonomous vehicle implementation predictions / implications for transport planning. *Victoria Transport Policy Institute*, 2025. URL https://www.vtpi.org/avip.pdf.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Richards, E. and Dai, L. Modeling imitation learning robustness to noisy demonstrations via sigmoid degradation. In

IEEE MIT Undergraduate Research Technology Conference (URTC), Cambridge, Massachusetts, 2025.

Richards, E., Lopez, A., Morales, J., and Zhang, Z. From chaos to clarity: Strengthening 3d collaborative autonomous vehicle perception with noise-aware training. In *IEEE MIT Undergraduate Research Technology Conference (URTC), Cambridge, Massachusetts*, 2025a. doi: TBD.

Richards, E., Thapa, B., and Mashayekhy, L. Edge-enabled collaborative object detection for real-time multi-vehicle perception. In *IEEE International Conference on Edge Computing and Communications (EDGE)*, pp. 13–22, 2025b. doi: 10.1109/EDGE67623.2025.00011.

Wali, B. The future on wheels: A joint heterogeneous model of automated vehicle adoption and perceived employment and economic impacts. *Transportation Research Part A: Policy and Practice*, 207:104941, 2026. ISSN 0965-8564. doi: https://doi.org/10.1016/j.tra.2026.104941. URL https://www.sciencedirect.com/science/article/pii/S0965856426000820.

Xia, Z., Chen, W., Wang, Y., and Yang, M.-H. Knowval: A knowledge-augmented and value-guided autonomous driving system. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3740–3749, June 2026.

Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., and Darrell, T. Bdd100k: A diverse driving dataset for heterogeneous multitask learning, 2020. URL https://arxiv.org/abs/1805.04687.

Zhang, Q., Wallbridge, C. D., Jones, D. M., and Morgan, P. L. Public perception of autonomous vehicle capability determines judgment of blame and trust in road traffic accidents. *Transportation Research Part A: Policy and Practice*, 179:103887, 2024. ISSN 0965-8564. doi: https://doi.org/10.1016/j.tra.2023.103887. URL https://www.sciencedirect.com/science/article/pii/S0965856423003075.

Figure 5. Top 4 most safe (top row) and hazardous (bottom row) images in the dataset, based on hazard score Λ.

A. Prompt sets and corruption ranges

Hazardous prompts: *a pedestrian crossing the road*, *a cyclist in the lane*, *a stopped vehicle blocking traffic*, *road debris on the street*, *a construction zone on the road*, *an emergency vehicle with flashing lights*. Non-hazardous prompts: *a clear road with no obstacles*, *a normal highway scene*, *an empty intersection*, *a clear lane ahead*, *a normal driving scene*. Corruption severities are uniformly discretized over the same ranges used in the class draft: fog density $[0.15, 0.75]$, Gaussian blur $\sigma \in [1, 8]$, motion-blur kernel size $[3, 25]$, JPEG quality $[90, 10]$, low-light scale $[0.8, 0.2]$, occlusion mask ratio $[0.05, 0.60]$, and downsampling scale $[0.90, 0.20]$.

B. Lambda value visualization

Fig. 5 shows the top 4 most and least hazardous scenes in the BDD100K dataset, according to our Λ function using the prompt sets listed in Appendix A. Notice how the non-hazardous images are mostly clear roads with minimal traffic, while the hazardous images depict more busy, complicated scenarios.

C. Corruption visualization

Fig. 6 illustrates each corruption operator being applied at various severity levels on real driving scenes. Each image is annotated with the corresponding hazard score (Λ), embedding drift (Δ), and margin drift (D_Λ).

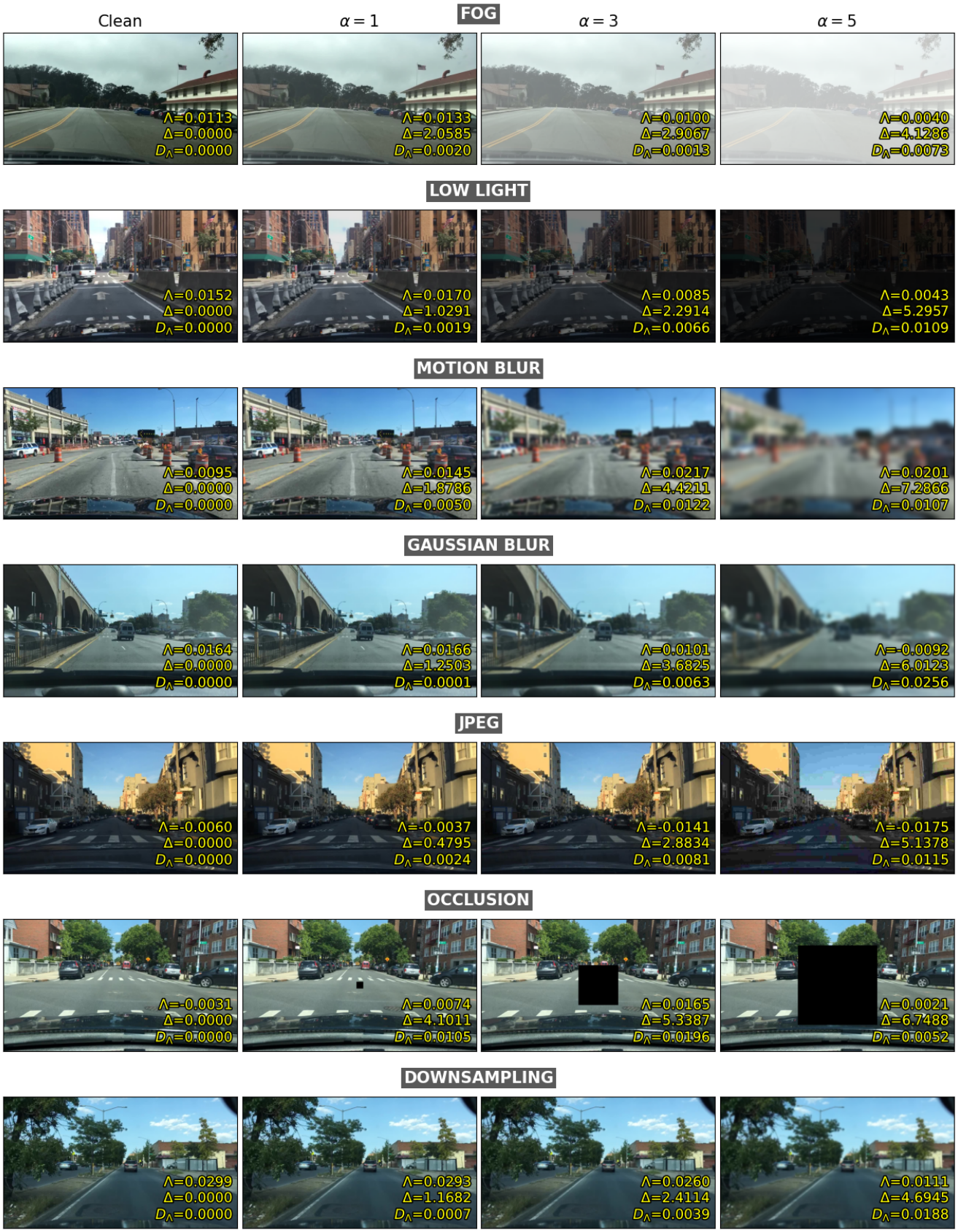


Figure 6. Visualizations of image corruptions on randomly-selected road images. Hazard level Λ, embedding drift Δ, and margin drift D_Λ are overlaid on each image.