The Sampling Cascade: How Data Collection Bias Creates Systematic Safety Blind Spots in VLM-Based Autonomous Driving

Xingnan Zhou and Ciprian Alecsandru

***Abstract*—Autonomous driving systems can only be as safe as the scenarios they have been trained on. Vision-language model (VLM) pipelines compound this limitation: a teacher model annotates collected data, and a student model learns from those annotations. We formalize this progressive diversity loss as the *sampling cascade*—a four-layer process through which scenario diversity narrows from real-world occurrence to model training. Defining a six-dimensional scenario space grounded in National Highway Traffic Safety Administration (NHTSA) pre-crash typologies and ISO 21448, we introduce the Safety Coverage Metric Φ for training-time combinatorial audit. Applied to the Waymo End-to-End dataset (WOD-E2E)—4,021 segments from 6.4 million miles—we find that 91% of compound scenario cells contain zero training examples and $\Phi < 1\%$. Scaling to 412K frames yields only marginal improvement, confirming the gap is structural. We trace downstream consequences through three analyses: Chain-of-Thought perception analysis reveals a 34.9% pedestrian miss rate with speed-behavior mismatches producing 40% higher trajectory error ($p < 0.001$); four safety-critical paradoxes emerge from 15 Bonferroni-corrected tests, including a vulnerable road user (VRU) visibility cliff (pedestrians drop 81% from day to night) and a yield sign paradox (3.5% of data concentrating 46% of VRU encounters); and a case study demonstrates that coverage-guided rebalancing enabled a 42-rank leaderboard improvement (rank 57 to 15 of 67). The framework is model- and dataset-agnostic, positioning Φ as a training-time complement to existing test-time safety validation.**

***Index Terms*—autonomous driving, dataset bias, vision-language models, vulnerable road users**

I. Introduction

AUTONOMOUS vehicles cannot yet demonstrate statistical safety parity with human drivers. Kalra and Paddock [1] estimate that proving a 20% fatality reduction would require 8.8 billion miles of on-road testing—a standard no current system approaches. The 2018 Uber ATG fatality [2] underscored the lethal consequences of encountering a scenario poorly represented in training: the system failed to classify a pedestrian crossing a dimly lit road. NHTSA statistics consistently identify nighttime conditions and vulnerable road users (VRUs) as overrepresented in fatal crashes [3], yet these are precisely the conditions that naturalistic data collection undersamples.

The AV industry has responded with increasingly sophisticated *test-time* scenario evaluation: Waymo's Collision Avoidance Testing, Aurora's Safety Case Framework, ISO 21448 (SOTIF) scenario validation [4]. A complementary question remains largely unexamined: does the *training data itself* cover the safety-critical scenario space?

VLM-based end-to-end (E2E) driving systems [5]–[7] have emerged as a dominant paradigm, using Chain-of-Thought (CoT) reasoning from large teacher models to train smaller student models via behavioral cloning. Systems such as CoVLA [8], DriveVLM [9], EMMA [10], and Senna [11] demonstrate the power of this approach but share a common vulnerability: training quality is bounded not only by teacher accuracy but also by the *representativeness of data presented to the teacher*. If the teacher never "sees" a cyclist in rain, the student cannot learn cyclist-in-rain avoidance regardless of its architecture.

This problem persists even in deliberately curated datasets. The Waymo Open Dataset for End-to-End Driving (WOD-E2E) [12]—mined from 6.4 million miles to capture 0.03%-frequency events, evaluated across 11 scenario clusters via human preference scores (RFS)—ensures each rare scenario *type* is represented. Yet type-level curation does not address *combinatorial* coverage: in the corpus we analyze, exactly 20 frames contain a cyclist at night, and zero contain a cyclist in rain.

In this paper, we formalize this phenomenon as the *sampling cascade*: a four-layer process through which the diversity of real-world driving scenarios is progressively narrowed before reaching the model's loss function:

1) **World** $\rightarrow$ **Collection**: Fleet driving patterns constrain observable scenarios.
2) **Collection** $\rightarrow$ **Selection**: Subset selection for annotation amplifies distributional bias.
3) **Selection** $\rightarrow$ **Annotation**: Teacher models accurately label the biased input.
4) **Annotation** $\rightarrow$ **Training**: Student models learn biased conditional distributions.

Crucially, no single layer is "wrong"—the gap emerges from their composition. We then trace the cascade's consequences through safety-critical findings, CoT perception analysis, and a model development case study.

Our contributions are:

- A **Sampling Cascade Framework** formalizing how data collection bias propagates through VLM training pipelines, creating compound safety gaps invisible to cluster-level curation.

Manuscript submitted March 2026.

X. Zhou and C. Alecsandru are with the Department of Building, Civil and Environmental Engineering, Concordia University, Montreal, QC H3G 1M8, Canada (e-mail: xingnan.zhou@mail.concordia.ca; ciprian.alecsandru@concordia.ca).

- A **Safety Coverage Metric** Φ, grounded in ISO 21448, for training-time combinatorial audit, shown to be robust across five weight configurations ($\Phi \in [0.67\%, 0.95\%]$).
- **Four safety-critical paradoxes** discovered through 15 Bonferroni-corrected statistical tests on 20,782 annotated frames, revealing how compound gaps create systematic failure modes.
- **CoT perception analysis** showing that speed-behavior mismatches in the student model's reasoning produce 40% higher trajectory error ($p < 0.001$), linking perceptual failures to kinematic outcomes.
- **Empirical validation** across 67 WOD-E2E leaderboard submissions showing that training coverage predicts per-cluster performance ($\rho = 0.71$, 92.5% of submissions positive), and a development case study where distributional analysis guided a 42-rank leaderboard improvement.

II. Related Work

A. End-to-End Autonomous Driving

VLM-based E2E driving [5]–[11], [13] has emerged as the dominant paradigm, using Chain-of-Thought (CoT) annotations from large teacher models for student training. CTL-Drive follows this paradigm, fine-tuning Qwen3-VL-4B [14] on teacher-annotated CoT labels via QLoRA. Poutine [15], the closest system to ours (ranked 3rd of 67 on WOD-E2E), shares the same VLM-teacher pipeline. The common reliance on teacher annotations creates a previously uncharacterized vulnerability: training quality is bounded not only by teacher accuracy but also by the representativeness of data presented to the teacher.

B. Safety Standards and Scenario-Based Validation

ISO 21448 (SOTIF) [4] mandates scenario-based testing to reduce the "unknown unsafe" region, and UL 4600 [16] extends this to full-lifecycle safety cases including data quality requirements; other frameworks (PEGASUS [17], Euro NCAP [18], ASAM OpenSCENARIO [19]) operationalize scenario-based evaluation at test time. All focus on *test-time* scenario coverage. Our contribution fills a gap: we analyze *training-time combinatorial* coverage using SOTIF's conceptual framework, mapping scenario cells to known/unknown and safe/unsafe quadrants based on training data presence rather than test results.

C. Dataset Bias in Autonomous Driving

Dataset bias in AV benchmarks is well-documented. Liu et al. [20] survey 70+ datasets and identify geographic, temporal, and weather biases as systemic. The Waymo Open Dataset [21] (primarily San Francisco and Phoenix) and nuScenes [22] (Boston and Singapore) each capture only a fraction of the global driving condition space. Feng et al. [23] quantify the long-tail distribution problem, showing that safety-critical scenarios follow power-law distributions with extreme rarity.

WOD-E2E [12] represents a major advance: segments are deliberately mined for long-tail events, and evaluation averages RFS across 11 scenario clusters to prevent models from gaming frequent scenarios. However, this type-level curation addresses which *categories* of rare events are present, not whether compound *combinations* of conditions are covered. To illustrate: WOD-E2E's Cyclist cluster ensures cyclist scenarios are evaluated, and its Construction cluster ensures construction-zone scenarios are evaluated, but neither addresses whether cyclist-at-construction-zone scenarios are covered in training. Our work shows that even within WOD-E2E's curated corpus, the combinatorial scenario space is 91% empty.

D. VRU Safety and Environmental Degradation

Vulnerable road users are overrepresented in AV-related incidents. NHTSA's FARS data [3] shows that 76% of pedestrian fatalities occur at night, and cyclists are 3.4× more likely to be killed at night than during the day. VRU detection degrades substantially under adverse conditions: Bijelic et al. [24] report 30–50% detection drops in fog; Zheng et al. [25] document rain degradation across multiple perception architectures. The VRU detection problem is particularly acute for VLM-based systems because the teacher model's VRU annotations determine the student's training signal: if the teacher cannot detect a pedestrian in fog, the student learns "no pedestrian present" for that scene.

These findings motivate our focus on VRU-weather-time compound scenarios. The sampling cascade ensures that the very conditions where VRU detection is most critical (night, rain) are also the conditions with the least training data—creating a systematic safety gap that no amount of model improvement can address without addressing the data.

III. The Sampling Cascade Framework

We formalize the progressive narrowing of scenario diversity from real-world occurrence to model training as the *sampling cascade*.

A. Scenario Space Definition

We define the driving scenario space $\mathcal{S}$ as the Cartesian product of six operationally relevant dimensions:

$$\mathcal{S} = \mathcal{T} \times \mathcal{W} \times \mathcal{V} \times \mathcal{I} \times \mathcal{C} \times \mathcal{U} \tag{1}$$

Table I details each dimension with its levels, cardinality, and safety rationale.

The dimensions capture principal factors from NHTSA pre-crash typologies [26] and SOTIF triggering conditions [4]. Each dimension uses coarse categorical levels rather than continuous variables, trading resolution for interpretability and statistical power.

B. Safety Criticality Weight

We define $w(s) \in [0, 1]$ as an additive, bounded function reflecting the safety criticality of each scenario cell:

$$w(s) = \min\big(w_{\text{VRU}}(v) + w_{\text{vis}}(t, w') + w_{\text{int}}(i) + w_{\text{tc}}(c) + w_{\text{spd}}(u),\ 1.0\big) \tag{2}$$

The component weights below are derived from NHTSA fatality proportions [3] and SOTIF risk categorizations [4], normalized

so that each dimension's maximum contribution reflects its share of crash severity in national statistics. The specific values serve as a baseline configuration; Section III-B's sensitivity analysis (Table II) confirms that all conclusions hold across five alternative weight schemes.

Component weights: **VRU** $\in [0, 0.35]$ (dominant, reflecting disproportionate fatality risk—NHTSA data shows VRUs account for 19% of all traffic fatalities [3]); **Visibility** $\in [0, 0.20]$ (night $+0.12$, rain $+0.08$, fog $+0.06$, snow $+0.10$, proportional to FARS nighttime and adverse-weather fatality rates); **Intersection** $\in [0, 0.15]$ (roundabout 0.12, Y 0.10, based on NHTSA pre-crash typology frequencies [26]); **Traffic control** $\in [0, 0.15]$ (yield 0.15 highest due to right-of-way ambiguity); **Speed** $\in [0, 0.15]$ (proportional to kinetic energy). The weight ranges from $w = 0.02$ (daytime, clear, no VRU, straight, uncontrolled, stopped) to $w = 0.97$ (nighttime, rain, both VRUs, roundabout, yield, fast), with mean $\bar{w} = 0.554$.

Worked Example. Consider the cell: *night, rain, cyclist, roundabout, yield, fast*. The component weights are: VRU(cyclist) $= 0.30$, visibility(night + rain) $= 0.12 + 0.08 = 0.20$, intersection(roundabout) $= 0.12$, traffic control(yield) $= 0.15$, speed(fast) $= 0.15$. Sum: 0.92, capped at $\min(0.92, 1.0) = 0.92$. This cell requires $n_{\text{req}} = 50 \times (1 + 3 \times 0.92) = 188$ training examples; it has zero.

Weight Sensitivity. To verify that our conclusions do not depend on the specific weight assignment, we evaluate Φ under five configurations: baseline (VRU-dominant), equal-weight, VRU-heavy, visibility-heavy, and intersection-heavy. Across all configurations, Φ remains below 1% (range: $[0.67\%, 0.95\%]$), and the fraction of Unknown Unsafe cells stays between 87–91% (Table II). The finding of massive coverage gaps is robust to weight specification. Even the most conservative (equal-weight) configuration, which minimizes the emphasis on VRU-heavy scenarios, still yields $\Phi = 0.95\%$—less than one-hundredth of the safety-critical scenario space is adequately covered.

C. Sampling Cascade Decomposition

The cascade describes progressive diversity reduction across four layers (Fig. 1).

Layer 1: World → Collection. WOD-E2E [12] is curated for long-tail events (6.4M miles → 0.03% frequency), ensuring *type-level* coverage of 11 scenario clusters. But fleet geography constrains *combinatorial* diversity: weather, time, and VRU co-occurrence patterns reflect collection cities (primarily San Francisco and Phoenix). Cyclists in rain, for instance, are rare not because the scenario is unimportant but because the collection geography does not capture it.

Layer 2: Collection → Selection. Of 415,663 extracted frames, 20,782 (5%) were selected for CoT annotation, stratified by intent (straight/left/right) and difficulty—not weather or VRU diversity. This selection amplifies the distributional bias from Layer 1.

Layer 3: Selection → Annotation. Two independent teacher models (Gemini Flash 3 [27] and Qwen3-VL-32B [14]) agree on >99.9% of binary VRU detections across 18,734 overlapping frames. The teacher introduces no additional bias; it accurately labels an unrepresentative input.

Layer 4: Annotation → Training. Zero cyclist-in-rain frames means $p(\texttt{cyclist} \mid \texttt{rain}) \approx 0$ in the student's learned distribution—correct for the training data, catastrophic for deployment. The student model cannot learn what the teacher never saw.

Cascade Properties. Two properties distinguish the sampling cascade from simple class imbalance: (1) *Monotonic loss*: each layer can only reduce or maintain diversity, never increase it. A diverse world can produce a biased collection, but a biased collection cannot produce diverse annotations. (2) *Composability*: the overall diversity loss is the product of per-layer losses, meaning that even small per-layer reductions compound into massive gaps. If each of four layers independently retains 50% of diversity, the cascade retains only $0.5^4 = 6.25\%$—and the actual retention is far worse because the layers are not independent but correlated (the same rare conditions are undersampled at every layer).

D. Safety Coverage Metric

We define the *Safety Coverage Metric* $\Phi \in [0, 1]$:

$$\Phi = \frac{\sum_{s \in \mathcal{S}} w(s) \cdot C(s)}{\sum_{s \in \mathcal{S}} w(s)} \tag{3}$$

where $C(s) = \min(n(s)/n_{\text{req}}(s), 1.0)$ and $n_{\text{req}}(s) = n_{\text{base}} \cdot (1 + 3 \cdot w(s))$ with $n_{\text{base}} = 50$, yielding $n_{\text{req}} \in [50, 200]$. The *Safety Coverage Gap* is $\Gamma = 1 - \Phi$. $\Phi = 1$ indicates that every cell has at least its required number of training examples; $\Phi = 0$ means no safety-critical cell has any coverage.

E. SOTIF Mapping

We map scenario cells to SOTIF quadrants [4] using observation status ($n(s) > 0$ = known) and risk level ($w(s) \geq 0.20$

TABLE I: Scenario Space Dimensions and Safety Rationale

Dim.	Levels	$\lvert\cdot\rvert$	Rationale
$\mathcal{T}$ Time	Day, Night, Dusk	3	NHTSA: 76% ped. fatalities at night
$\mathcal{W}$ Weather	Clear, Fog, Rain, Snow	4	SOTIF triggering condition
$\mathcal{V}$ VRU	None, Ped, Cyclist, Both	4	Highest fatality risk
$\mathcal{I}$ Intersection	None, Cross, T, Y, Merge, Roundabout	6	NHTSA pre-crash typology
$\mathcal{C}$ Traffic ctrl.	None, Green, Red, Stop, Yield	5	Right-of-way complexity
$\mathcal{U}$ Speed	Stopped, Slow, Moderate, Fast	4	Kinetic energy / stopping distance
Total cells		**5,760**	$3\times4\times4\times6\times5\times4$

TABLE II: Weight Sensitivity Analysis

Configuration	Φ (%)	Γ (%)	Unk. Unsafe
Baseline (VRU-dominant)	0.75	99.25	91.0%
Equal weight	0.95	99.05	87.0%
VRU-maximized	0.67	99.33	88.6%
Visibility-maximized	0.78	99.22	90.8%
Intersection-maximized	0.80	99.20	91.0%

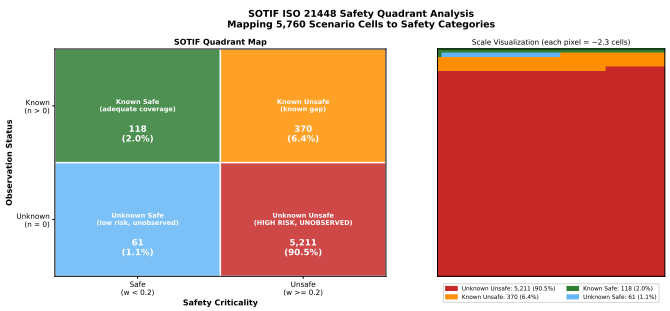

Fig. 1: SOTIF quadrant mapping of the 5,760 scenario cells. The Unknown Unsafe region (90.5%) dwarfs all other quadrants. Known Unsafe (6.4%) represents observed but insufficiently covered high-risk cells. Only 2.0% of cells are Known Safe.

= unsafe): **Known Safe**—low risk, observed, or high risk with $C \geq 0.5$; **Known Unsafe**—high risk, observed, $C < 0.5$; **Unknown Unsafe**—high risk, $n = 0$ (most dangerous); **Unknown Safe**—low risk, $n = 0$. The distribution (Fig. 1) reveals that 90.5% of cells are Unknown Unsafe—the scenario category that SOTIF considers the highest priority for reduction.

IV. Coverage Analysis

A. Dataset and Annotation Pipeline

WOD-E2E [12] contains 4,021 segments (2,037 training) curated for long-tail scenarios from 6.4M miles. Performance is evaluated via the Rater Feedback Score (RFS)—human preference ratings averaged equally across 11 scenario clusters—with ADE as tiebreaker. This cluster-balanced evaluation prevents gaming frequent scenarios; our analysis asks whether compound combinations within and across clusters are also covered.

From the 415,663 extracted training frames, 20,782 were annotated using Gemini Flash 3 [27] with binary presence flags for 12 object categories, natural language explanations, and speed/command labels. Intersection type, traffic control, and weather are derived via rule-based extraction from explanations, cross-validated against WayGraph intersection fingerprints [28]. The full 412K frames were additionally annotated using Qwen3-VL-32B [14]; due to its North American signage recognition limitations (Section IV-C), the 412K analysis uses a reduced 5D space (omitting traffic control). The student model is CTL-Drive: Qwen3-VL-4B + QLoRA ($r = 128$, $\alpha = 256$), ranked 15th of 67 on the WOD-E2E leaderboard (RFS = 7.705).

B. Coverage Results

The vast majority of the scenario space is empty: over 91% of cells contain zero training examples (Fig. 2). The safety-weighted coverage is $\Phi = 0.75\%$—meaning that for every 100 units of safety-critical training need, the corpus provides less than one. The gap is worst precisely where it matters most: snow scenarios have 460× less coverage than clear weather; cyclists 16× less than vehicle-only scenes; roundabouts 235× less than standard crossings. Cells with the highest safety weight consistently exhibit the lowest coverage.

TABLE III: Scenario Space Summary and Scale Validation

Metric	F3 20K (6D)	F3 20K (5D)	Q32B 412K (5D)
Frames	20,782	20,782	412,174
Total cells	5,760	1,152	1,152
Occupied	512 (8.9%)	148 (12.8%)	241 (20.9%)
Φ	0.75%	2.5%	6.0%
Γ	99.25%	97.5%	94.0%
Unknown Unsafe	5,211 (90.5%)	979 (85.0%)	895 (77.7%)
Φ gain	—	—	2.4×

Scale Validation (Φ Scaling). On 412,174 Qwen3-VL-32B annotations (5D space, 1,152 cells): Flash 3 (20K) covers 148 cells ($\Phi_{5D} = 2.5\%$); Qwen 32B (412K) covers 241 ($\Phi_{5D} = 6.0\%$). A 20× data increase yields only 2.4× Φ improvement; 77.7% of cells remain Unknown Unsafe (Table III). The sub-linear scaling confirms compound gaps are *structural*, not a sample-size problem—they arise from the geographic and temporal constraints of naturalistic data collection. Along every dimension, the most safety-critical levels (night, snow, cyclist, roundabout, yield) have the lowest coverage. The top-10 critical empty cells all involve nighttime adverse weather with both VRU types at complex intersections; each requires 186–195 training examples but has zero.

C. Inter-Teacher Agreement

Two independent teacher models (Gemini Flash 3 and Qwen3-VL-32B) were compared on 18,734 overlapping frames. Binary VRU detection shows near-perfect agreement: >99.9% for pedestrians and cyclists across all conditions, including nighttime ($n = 5{,}207$) and night+rain ($n = 795$) subsets ($\kappa \geq 0.989$). This confirms that VRU absence reflects genuine absence in the collected images, not teacher limitations. Fine-grained signage classification diverges (overall 77.5% agreement; yield sign precision 0.095 due to 1,664 Qwen 32B false positives), motivating our use of Flash 3 for traffic control and restricting the 412K Qwen 32B corpus to a 5D space omitting traffic control.

V. Safety-Critical Findings

We conducted 15 pre-specified statistical tests on the 20,782 annotated frames, applying Bonferroni correction ($\alpha = 0.05/15 = 0.0033$). Fourteen of 15 tests achieve significance (Table IV). Four findings constitute novel safety-critical paradoxes.

A. The VRU Visibility Cliff

VRU annotations collapse catastrophically under adverse conditions. During daytime ($n = 14{,}654$), 24.5% of frames contain a VRU; at night ($n = 5{,}881$), this drops to 4.5%—a 5.5-fold reduction. Disaggregated by VRU type:

- **Pedestrians:** 21.7% (day) → 4.1% (night); odds ratio (OR) = 6.41, $p < 10^{-250}$.
- **Cyclists:** 5.3% (day) → 0.34% (night); OR = 16.43, $p < 10^{-87}$.

The cyclist odds ratio is 2.6× that of pedestrians, indicating two-wheeled VRUs are disproportionately lost in low-light

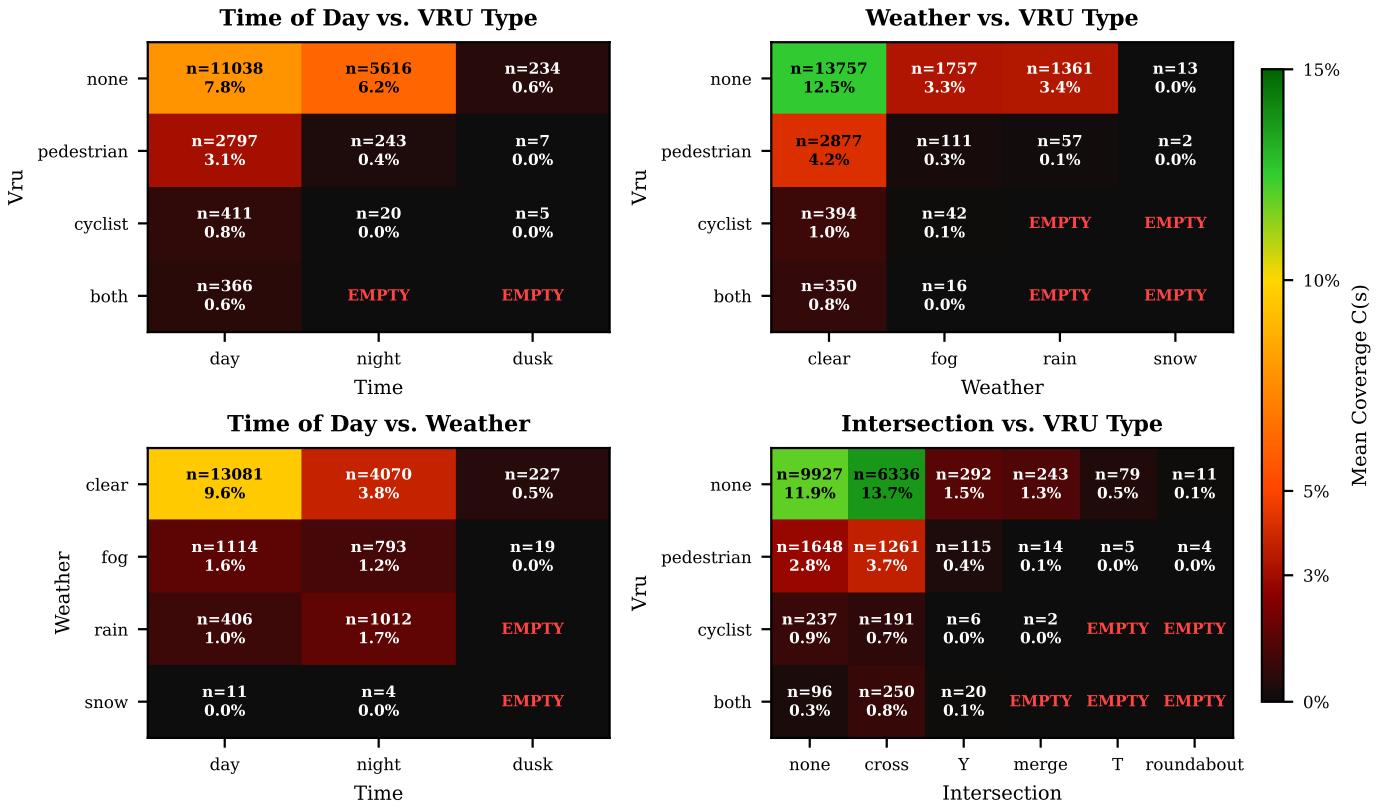

Fig. 2: Scenario Space Coverage: 2D projections of the 6D scenario space. Each cell shows the fraction of the required training samples present. White cells indicate zero observations. The majority of the space—particularly combinations involving adverse weather, VRU presence, and complex intersections—is completely unobserved.

conditions. Only 20 cyclist annotations exist across all 5,881 nighttime frames—a rate so low that any model trained on these labels would have no statistical basis for learning cyclist avoidance at night.

Under adverse weather (rain+fog+snow, $n = 3{,}360$), VRU rate drops to 6.8% vs. 20.9% in clear conditions (OR $= 3.60$, $p < 10^{-97}$). Rain ($n = 1{,}418$) contains **zero cyclist annotations**—a complete absence. Under compound conditions (night+rain), 95.6% of expected VRU observations are eliminated (Fig. 3).

This pattern mirrors the conditions of the 2018 Uber ATG fatality [2]: a nighttime pedestrian crossing comprising less than 0.1% of this training corpus.

B. The Yield Sign Paradox

Yield sign scenarios constitute only 3.5% of the dataset ($n = 730$) but exhibit the highest VRU rate (46.0%) and conflict rate (41.0%) of any traffic control category—a 2.6× and 4.5× multiplier over the dataset average, respectively. A model trained with uniform loss weighting encounters yield sign scenarios in only 1 out of every 28 gradient updates, yet these scenarios demand the most complex decision-making: simultaneously tracking VRUs, resolving right-of-way ambiguity, and negotiating with conflicting vehicles.

Cross-intersections ($n = 8{,}062$) are the most common intersection type with VRUs, but Y-intersections ($n = 433$), despite being rare, exhibit the highest pedestrian prevalence at

Fig. 2. VRU annotation cascade loss under adverse conditions

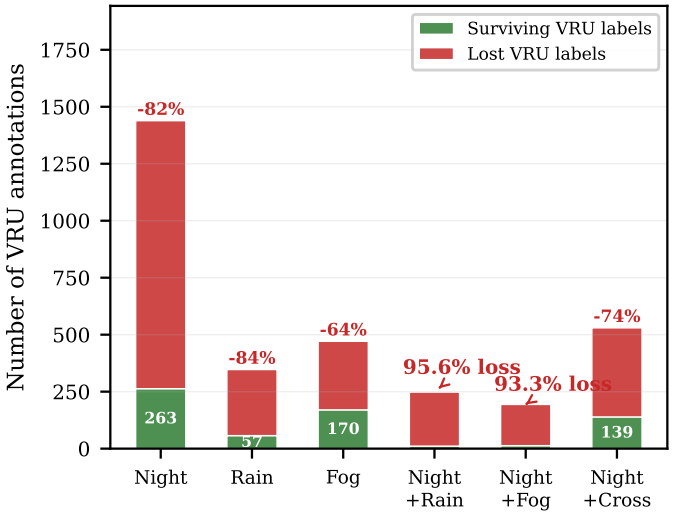

Fig. 3: The VRU cascade: progressive loss of VRU annotations as conditions worsen. Starting from the full corpus, each adverse condition (night, rain, fog) removes a further fraction of VRU observations. The compound effect of night+rain eliminates 95.6% of expected VRU training signal.

31.2% (Kruskal-Wallis $H = 156.1$, $p < 10^{-31}$). The training distribution is inversely correlated with safety criticality.

C. The Green Light Paradox

Green traffic signals contain a higher VRU rate (20.7%) than red light scenes (16.4%; OR $= 1.34$, $p = 0.022$), yet the teacher model's annotated deceleration rate at green lights is only 6.8% compared to 38.2% at red lights—a 5.6× differential. This creates a systematic signal-action conflation: the model learns that green licenses acceleration, but the actual safety-critical variable is VRU presence, not signal state.

The paradox is ecologically valid: pedestrians legitimately cross at signalized intersections during green phases (crosswalk signals are concurrent with through-movement green), and cyclists share the road. End-to-end models that learn from behavioral cloning inherit this conflation because ground-truth ego trajectories at green lights are predominantly maintain-speed or accelerate, regardless of crosswalk activity.

D. Survivorship Bias in Nighttime VRU Response

Among the 263 frames where VRUs are detected at night, the deceleration rate is 31.2%—significantly *higher* than the daytime VRU deceleration rate of 20.6% (OR $= 0.57$, $p = 0.002$). This counter-intuitive result reflects survivorship bias: VRUs that cross the nighttime detection threshold are disproportionately close, salient, and collision-imminent. Distant or peripheral VRUs, which during daytime would be annotated as "present" without requiring speed adjustment, are simply not detected at night.

The dangerous implication: performance metrics computed only over detected-VRU frames will overstate nighttime safety, because they exclude the most dangerous cases—undetected VRUs—by construction. The model achieves acceptable braking on the *visible* subset while potentially failing on the *invisible* majority.

E. Pedestrian vs. Cyclist Profile Divergence

Pedestrians and cyclists demand fundamentally different safety strategies. Cyclist encounters occur at 87% higher ego speed (median 6.81 vs. 3.64 m/s; $r = 0.30$, $p < 10^{-39}$) and predominantly involve in-road sharing (63.1%) rather than crossing (34.0%). The 4.3:1 pedestrian-to-cyclist training ratio may prevent the model from learning adequate cyclist-specific behaviors, particularly given the near-total nighttime cyclist erasure documented above.

VI. CoT Perception and Trajectory Analysis

The coverage gaps identified above affect the model's training distribution. We now examine whether these gaps manifest in the student model's Chain-of-Thought reasoning and trajectory predictions.

A. CoT Perception Accuracy

We evaluated CTL-Drive's CoT predictions on 472 frames with valid JSON output against Gemini Flash 3 teacher annotations (Table V). The model accurately detects common objects—nearby vehicles (F1 = 0.97), traffic elements (F1 = 0.93), construction zones (F1 = 0.89)—but is not as successful with safety-critical categories:

TABLE IV: Summary of 15 Statistical Tests (Bonferroni-Corrected)

ID	Comparison	Effect	Adj. p	Sig.
T1	VRU day vs. night	OR = 6.92	$< 10^{-295}$	Yes
T2	Ped day vs. night	OR = 6.41	$< 10^{-250}$	Yes
T3	Cyclist day vs. night	OR = 16.43	$< 10^{-87}$	Yes
T4	VRU clear vs. adverse	OR = 3.60	$< 10^{-97}$	Yes
T5	Speed: VRU vs. non-VRU	$\chi^2 = 78.5$	$< 10^{-16}$	Yes
T6	Speed: ped vs. cyclist	$\chi^2 = 32.8$	$< 10^{-6}$	Yes
T7	Ego spd: VRU vs. non	$r = 0.24$	$< 10^{-120}$	Yes
T8	Ego spd: ped vs. cyclist	$r = 0.30$	$< 10^{-39}$	Yes
T9	VRU by intersection	$H = 156.1$	$< 10^{-31}$	Yes
T10	Conflict: cross vs. none	OR = 3.35	$< 10^{-144}$	Yes
T11	Decel: VRU+day vs. night	OR = 0.57	.002	Yes
T12	VRU: green vs. red	OR = 1.34	.022	Yes
T13	Conflict: day vs. night	OR = 1.31	$< 10^{-6}$	Yes
T14	Cyclist: clear vs. fog	OR = 1.44	.101	No
T15	Speed: red vs. stop	$\chi^2 = 139.1$	$< 10^{-30}$	Yes

TABLE V: CoT Perception Accuracy (CTL-Drive V8, $n = 472$)

Category	Prec.	Rec.	F1	TP	FP	FN
Nearby vehicle	.959	.984	.971	422	18	7
Traffic element	.933	.936	.934	320	23	22
Cyclist	.960	.910	.934	71	3	7
Construction	.919	.862	.890	125	11	20
Weather cond.	.747	.849	.795	62	21	11
Pedestrian	**.773**	**.651**	**.707**	**99**	**29**	**53**
Special vehicle	.707	.558	.624	29	12	23
Confl. veh.	**.705**	**.449**	**.549**	**31**	**13**	**38**

- **Pedestrians:** 34.9% miss rate (53/152 ground-truth positives missed), F1 = 0.71. The model fails to detect one in three pedestrians flagged by the teacher.
- **Conflicting vehicles:** 55.1% miss rate (38/69), F1 = 0.55. Right-of-way conflicts are missed more often than detected.
- **Cyclists:** 9.0% miss rate (7/78), F1 = 0.93. Substantially better than pedestrians, possibly because cyclist visual signatures are more distinctive.

The frame classification reveals: 33.1% Confident Correct (CC), 57.0% Confident Wrong (CW), 7.4% Hedged Cautious (HC), and 2.5% Hedged Wrong (HW). The high CW rate (model confidently produces incorrect object detections) is particularly concerning for safety: these frames provide no epistemic signal to downstream safety monitors.

Stratified analysis. Nighttime dangerous misses (speed under-reaction when deceleration was needed) occur at 16.5% vs. 11.7% during daytime. Yield sign scenarios show the highest hedging rate (40.0%), consistent with the Yield Sign Paradox's identification of these scenarios as the most complex. Hard scenarios (by trajectory difficulty) show nearly $4\times$ the hedging rate of easy scenarios (19.4% vs. 4.9%), suggesting the model's epistemic uncertainty is appropriately calibrated to difficulty even if its object detection is not.

Speed behavior analysis. Of 472 valid frames, 75.0% show speed-behavior matches between model and teacher, while 12.5% are dangerous misses (model recommends maintain/accelerate when teacher says decelerate). The dangerous miss rate is highest for yield signs (40.0%) and merge intersections (27.3%)—again the most safety-critical categories.

TABLE VI: Trajectory Error by CoT Frame Class ($n = 472$)

Frame Class	n	ADE	FDE	Unsafe	Note
Conf. Correct (CC)	156	1.36	3.03	5.1%	Best
Conf. Wrong (CW)	269	1.41	3.10	5.9%	Highest unsafe
Hedged Wrong (HW)	12	0.74	1.49	0.0%	Cautious
Hedged Cautious (HC)	35	0.90	2.08	5.7%	
Speed: match	354	1.23	—	—	
Speed: dangerous miss	59	1.72	—	—	$p < .001$

B. CoT–Trajectory Correlation

We matched CoT frame classifications to trajectory metrics (ADE/FDE) on the same 472 frames to test whether reasoning quality predicts trajectory accuracy (Table VI).

Across the four frame classes, a Kruskal-Wallis test reveals significant ADE differences ($H = 16.5$, $p < 0.001$). The most striking result is that **speed-behavior mismatches produce 40% higher trajectory error**: frames where the model's speed recommendation is a "dangerous miss" (recommending maintain/accelerate when the teacher annotated decelerate) have ADE $= 1.72$ m vs. 1.23 m for speed matches ($p < 0.001$).

Hedged frames (HC + HW, $n = 47$) show *lower* ADE than confident frames (0.87 vs. 1.40 m). This counter-intuitive finding suggests that the model's hedging language correlates with cautious trajectories—uncertainty in reasoning produces conservative kinematics, which happens to be safer. The implication is that *confident wrong* (CW) frames are more dangerous than *hedged wrong* (HW) frames: confident errors provide no warning signal.

VII. Case Study: Distributional Analysis in Practice

We illustrate how the sampling cascade framework guided concrete model improvements by documenting the development of CTL-Drive from V4 (RFS 6.538, rank 57/67) to V8 (RFS 7.705, rank 15/67)—a 42-rank improvement on the WOD-E2E leaderboard.

A. Diagnosis via Coverage Analysis

Applying the cascade framework to our training data revealed two actionable distributional imbalances:

Turn-ratio imbalance. The training distribution was heavily skewed: through movements 84%, left turns 8.6%, right turns 7.2%. Given the trajectory validation finding that right turns produce 62% higher error than left turns (Section IV), this imbalance was directly impacting performance on turn-heavy evaluation clusters.

Loss masking bug. V4 used the default messages format for training, which applied loss on the full conversation including the prompt template. V8 fixed this with proper prompt/completion masking, ensuring gradient signal was concentrated on the model's actual output (CoT + trajectory).

B. Treatment: Targeted Rebalancing

Based on the coverage analysis, V8 applied two targeted interventions:

TABLE VII: Training Configuration: V4 vs. V8

Parameter	V4	V8
Base model	Qwen3-VL-4B	Qwen3-VL-4B
LoRA (r, α)	128, 256	128, 256
CoVLA pre-train	Stage 1a V4	Stage 1a V8
Loss masking	Messages (bug)	Prompt/completion
Turn rebalancing	None	5×L, 4×R
Effective samples	20,782	90,079
Training epochs	2	2
Training time	∼7h	∼14.5h
Camera input	Front only	Front only
RFS (Overall)	**6.538**	**7.705**
Rank	**57/67**	**15/67**

Turn rebalancing. Left turns were oversampled $5\times$ and right turns $4\times$, expanding the effective training set from 20K to 90K frames. The asymmetric oversampling (more for left than right) reflects the observation that left turns, while slightly more frequent in the data (8.6% vs. 7.2%), involve longer exposure in opposing traffic lanes and higher collision risk. After rebalancing, the turn distribution was approximately: through movements 62%, left turns 21%, right turns 17%.

Loss masking fix. V4 used the default messages format for training, which applied loss to the full conversation including the system prompt and user template tokens. This diluted gradient signal: of the $\sim$2,000 tokens per training example, only $\sim$500 (the CoT output and trajectory) were informative, but V4 computed loss over all $\sim$2,000. V8 switched to proper prompt/completion masking, concentrating gradient signal on the model's actual output. The training loss dropped from 0.985 (V8 initial) to 0.635 (final) over 22,520 steps (14.5 hours on a single RTX 4090).

Table VII summarizes the configuration changes.

C. Validation: Per-Cluster Impact

The RFS improvement of 1.167 points was not uniform across clusters. Turn-heavy clusters showed the largest gains, consistent with the targeted rebalancing. CTL-Drive V8's per-cluster RFS ranges from 6.658 (Spotlight—the most extreme long-tail cluster) to 8.064 (Single-lane), a 1.4-point spread that mirrors the training coverage gradient. Notably, V8 achieves RFS 8.027 on Special Vehicles—*exceeding* the #1-ranked model NTR (7.751) on this cluster—demonstrating that targeted distributional analysis can produce cluster-specific improvements that outperform overall leaders.

The V4→V8 improvement is attributable to both interventions. To estimate their relative contributions, we note that the loss masking fix affects all training examples equally (improving gradient quality), while turn rebalancing specifically targets turn-related clusters. The fact that non-turn clusters (Construction, FOD) also improved suggests the loss masking fix contributed substantially. A controlled ablation (V8 without turn rebalancing) was not conducted due to computational constraints but would isolate the individual effects.

Oversampling can only rebalance categories where training examples *exist*. For the 91% of compound cells with zero frames, no resampling helps—these require simulation (e.g., CARLA [29]) or dedicated collection, prioritized by $w(s)$.

VIII. Discussion

A. Coverage-Performance Link Across 67 Models

To test whether the coverage-performance relationship generalizes beyond CTL-Drive, we analyzed all 67 WOD-E2E leaderboard submissions [30]. Mapping 8 of the 11 evaluation clusters to training-data proxies derived from our annotations (excluding FOD, Spotlight, and Others, which lack direct annotation equivalents), the median per-cluster RFS correlates with training data volume (Spearman $\rho = 0.71$, $p = 0.047$; Fig. 4). Clusters with fewer training examples—Cyclist (803 frames, median RFS 7.35) and Multi-lane (811 frames, 7.22)—consistently produce the lowest RFS; well-represented clusters such as Construction (4,416 frames, 7.95) and Single-lane (6,712 frames, 7.90) yield the highest.

This pattern holds for 92.5% of individual submissions (positive per-submission ρ), with median per-submission $\rho = 0.67$, confirming that the coverage-performance link is architecture-independent. RFS and ADE@5s are strongly correlated across 67 submissions ($\rho = -0.78$, $p < 10^{-4}$), validating ADE as a reasonable proxy, though not a substitute, for human preference evaluation [12].

B. Compound Interactions

The four safety-critical paradoxes (Section V) reveal compound interactions invisible to one-dimensional evaluation: **Multiplicative erasure**—night $\times$ rain eliminates 95.6% of VRU observations, worse than the $19.8\times$ expected under independence. **Inverse frequency-criticality**—yield signs (3.5% of data) concentrate 46% of VRU encounters, inverting the frequency-criticality relationship that uniform loss weighting assumes. **Signal-action conflation**—green lights suppress VRU-appropriate caution despite higher VRU rates than red lights, because behavioral cloning inherits the signal→accelerate association from ground-truth trajectories. **Survivorship bias**—nighttime VRU deceleration appears higher than daytime only because detected nighttime VRUs are disproportionately close and salient, systematically overstating nighttime safety. These compound interactions explain why cluster-level evaluation provides necessary but insufficient safety assurance: the cross-product of conditions falls through the evaluation cracks.

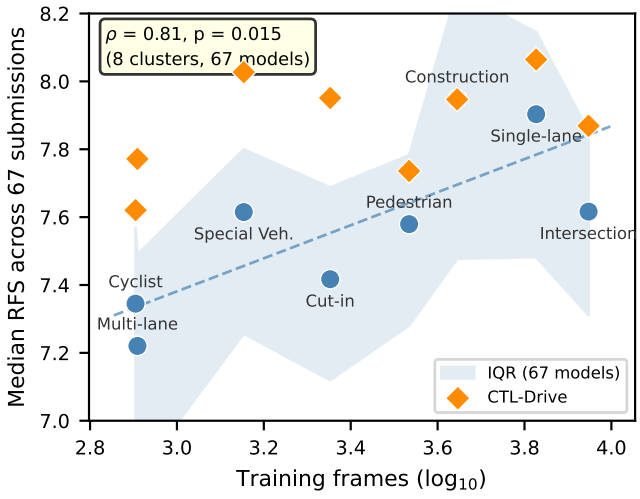

Fig. 4: Training data coverage predicts leaderboard performance. Each point is one of 8 WOD-E2E scenario clusters; blue circles show median RFS across all 67 submissions, orange diamonds show CTL-Drive. Clusters with fewer training examples produce systematically lower RFS.

C. Toward Solutions

The framework prescribes specific interventions for each gap type:

Sparse cells ($n > 0$): Targeted oversampling, as demonstrated in our V4→V8 case study (Section VII). The 42-rank improvement validates this approach for cells with some coverage. The key insight is that oversampling should be guided by Φ, not just by class frequency—a cell with $n = 5$ and $w = 0.9$ (five nighttime cyclist encounters) needs more resampling than a cell with $n = 5$ and $w = 0.1$ (five daytime vehicle-only scenes).

Empty cells ($n = 0$): Simulation (CARLA [29], ASAM OpenSCENARIO [19]), synthetic generation, or dedicated collection campaigns. The top-10 critical empty cells should be prioritized: all involve nighttime adverse weather with both VRU types at complex intersections. Simulation is particularly suitable because these compound conditions are individually well-understood (rain rendering, nighttime lighting, VRU animation) even if their combination is rarely observed in nature.

Signal-action decoupling: The Green Light Paradox suggests restructuring CoT reasoning to require separate "signal state" and "VRU proximity" steps before trajectory generation. Rule-based safety constraints (e.g., RSS [31]) can provide a safety floor independent of learned associations.

Standards extension: SOTIF [4] could be extended to require combinatorial Φ reporting alongside test-time scenario coverage mandates. UL 4600 [16] already requires data quality documentation; Φ provides a quantitative metric for this purpose.

D. Trajectory Validation

The coverage analysis establishes structural gaps; trajectory analysis confirms that these gaps manifest in model performance. On 2,000 stratified frames from 73 scenario cells, CTL-Drive achieves ADE = 1.51 m and FDE = 3.42 m overall, with 5.1% of frames producing unsafe trajectories (FDE > 8 m). The stratified breakdown (Table VIII) reveals three key patterns:

More complex intersections produce worse predictions. Error increases monotonically from straight roads (1.38 m) through standard crossings (1.62 m) to merge points (2.10 m), a significant gradient (Kruskal-Wallis $p = 0.004$).

Right turns are substantially harder than left turns. Right-turn error is 62% higher (1.93 vs. 1.19 m; $p < 0.001$), with $9\times$ the unsafe rate, despite similar training proportions.

VRU presence does not affect trajectory quality—an informative null ($p = 0.245$), suggesting the cascade's greatest danger is perceptual rather than kinematic: the model produces equally good *trajectories* whether VRUs are present or not, but may not *perceive* VRUs correctly.

TABLE VIII: Trajectory Validation by Scenario (CTL-Drive V8)

Scenario	n	ADE	FDE	Unsafe	p
Overall	2,000	1.508	3.417	5.1%	—
Intersection (Kruskal-Wallis)					**.004**
none	1,087	1.384	3.112	5.1%	
cross	729	1.621	3.703	4.9%	
merge	50	2.101	4.872	6.0%	
multi	8	3.176	8.173	12.5%	
Command (Mann-Whitney)					$<$ **.001**
straight	1,749	1.493	3.399	5.0%	
right turn	143	1.927	4.309	8.4%	
left turn	108	1.186	2.519	0.9%	
VRU (Mann-Whitney)					.245
present	638	1.470	3.290	4.6%	
absent	1,362	1.525	3.476	5.3%	
Difficulty (Mann-Whitney)					$<$ **.001**
easy	922	1.396	3.138	3.0%	
hard	501	1.845	4.188	9.0%	

E. Limitations

The six-dimensional scenario space omits variables such as road curvature and traffic density; adding dimensions would worsen coverage gaps. Safety weights are derived from NHTSA fatality data rather than calibrated to a specific crash database, though the sensitivity analysis (Table II) demonstrates robustness. Trajectory validation uses 2,000 frames with small samples for rare categories (multi: $n = 8$); claims for these groups are exploratory. The CoT analysis is limited to 472 frames due to a 76.3% JSON parse failure rate, which itself suggests the model's structured output generation is not robust. CTL-Drive's CoVLA pre-training ($\sim$300K frames [8]) may provide some compound scenario coverage that our fine-tuning-only analysis does not capture. The coverage-RFS correlation ($\rho = 0.71$) uses approximate cluster-to-annotation mappings for 8 of 11 clusters.

IX. Conclusion

Data-driven driving systems can only be as safe as the scenarios they train on. We have shown that even WOD-E2E—deliberately curated for long-tail scenarios from 6.4 million miles—covers less than 1% of safety-critical *compound* scenario combinations, and that this gap is structural: increasing the dataset twentyfold barely reduces it. The Safety Coverage Metric Φ remains below 1% regardless of weight configuration, and 90.5% of cells are classified as SOTIF Unknown Unsafe.

The cascade's downstream consequences are concrete: a 34.9% pedestrian miss rate in CoT perception, speed-behavior mismatches producing 40% higher trajectory error, a VRU visibility cliff that eliminates 95.6% of nighttime rain VRU observations, and a yield sign paradox concentrating 46% of VRU encounters in 3.5% of the data. The coverage-performance link holds across 67 leaderboard submissions ($\rho = 0.71$), confirming that the gap is architecture-independent.

The root cause is not annotation quality, model capacity, or insufficient curation. Two independent teachers agree on what they see; the problem is that compound safety-critical conditions are inherently undersampled by naturalistic collection. The framework prescribes concrete next steps: (1) training-time Φ audit before model training; (2) targeted oversampling for sparse cells, as demonstrated by our 42-rank improvement; (3) simulation for the 91% of empty cells; and (4) extending SOTIF and UL 4600 to require combinatorial Φ reporting. The sampling cascade cannot be broken after the fact; it must be addressed at its source.

References

[1] N. Kalra and S. M. Paddock, "Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability?" *Transportation Research Part A: Policy and Practice*, vol. 94, pp. 182–193, 2016.

[2] National Transportation Safety Board, "Collision between vehicle controlled by developmental automated driving system and pedestrian, Tempe, Arizona, March 18, 2018," National Transportation Safety Board, Tech. Rep. NTSB/HAR-19/03, 2019, highway Accident Report.

[3] National Highway Traffic Safety Administration, "Fatality analysis reporting system (FARS): 2023 annual report file," U.S. Department of Transportation, Tech. Rep., 2023, 76% of pedestrian fatalities occur at night; VRUs account for 19% of all traffic fatalities.

[4] International Organization for Standardization, "ISO 21448:2022 — road vehicles — safety of the intended functionality," International Standard, 2022, geneva, Switzerland.

[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)*, June 2023, pp. 17 853–17 862.

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

[7] X. Weng, B. Ivanovic, Y. Wang, Y. Wang, and M. Pavone, "PARA-Drive: Parallelized architecture for real-time autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 15 449–15 458.

[8] H. Arai, K. Miwa, K. Sasaki, Y. Yamaguchi, K. Watanabe, S. Aoki, and I. Yamamoto, "CoVLA: Comprehensive vision-language-action dataset for autonomous driving," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025, arXiv preprint arXiv:2408.10845, 2024.

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

[10] J.-J. Hwang, R. Xu, H. Lin, W.-C. Hung, J. Ji, K. Choi, D. Huang, T. He, P. Covington, B. Sapp, Y. Zhou, J. Guo, D. Anguelov, and M. Tan, "EMMA: End-to-end multimodal model for autonomous driving," 2024.

[11] 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," 2024.

[12] R. Xu, H. Lin, W. Jeon, H. Feng, Y. Zou, L. Sun, J. Gorman, E. Tolstaya, S. Tang, B. White, B. Sapp, M. Tan, J.-J. Hwang, and D. Anguelov, "WOD-E2E: Waymo Open Dataset for end-to-end driving in challenging long-tail scenarios," 2025, 4,021 segments curated from 6.4M miles; RFS evaluation across 11 scenario clusters.

[13] M. Nie, R. Peng, C. Wang, X. Cai, J. Han, H. Xu, and Z. Zhang, "Reason2Drive: Towards interpretable and chain-based reasoning for autonomous driving," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

[14] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin, "Qwen2.5-VL technical report," 2025, qwen3-VL models follow the same architecture.

[15] L. Rowe, R. de Schaetzen, R. Girgis, C. Pal, and L. Paull, "Poutine: Vision-language-trajectory pre-training and reinforcement learning post-training enable robust end-to-end autonomous driving," 2025, mila – Québec AI Institute / Université de Montréal.

[16] Underwriters Laboratories, "UL 4600: Standard for safety for the evaluation of autonomous products," Safety Standard, Edition 1, 2020, covers safety case framework, data quality, and lifecycle management for autonomous systems.

[17] H. Winner, K. Lemmer, T. Form, and J. Mazzega, "PEGASUS — first steps for the safe introduction of automated driving," German Federal Ministry for Economic Affairs and Energy (BMWi), Tech. Rep., 2019, project for the Establishment of Generally Accepted Quality Criteria, Tools and Methods as well as Scenarios and Situations for the Release of Highly Automated Driving Functions. See also Springer Lecture Notes in Mobility, https://doi.org/10.1007/978-3-319-94896-6_16.
[18] European New Car Assessment Programme (Euro NCAP), "Assessment protocol — assisted driving: Highway assist systems," Test and Assessment Protocol, v1.1, 2023, available: https://www.euroncap.com/en/for-engineers/protocols/.
[19] Association for Standardization of Automation and Measuring Systems (ASAM), "OpenSCENARIO: Dynamic content in driving simulation," Standard, v1.0, 2020, available: https://www.asam.net/standards/detail/openscenario/.
[20] M. Liu, E. Yurtsever, J. Fossaert, X. Zhou, W. Zimmer, Y. Cui, B. L. Zagar, and A. C. Knoll, "A survey on autonomous driving datasets: Statistics, annotation quality, and a future outlook," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 11, pp. 7138–7164, 2024.
[21] 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, Y. Zhang, J. Shlens, Z. Chen, and D. Anguelov, "Scalability in perception for autonomous driving: Waymo Open Dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020, pp. 2443–2451.
[22] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuScenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 621–11 631.
[23] S. Feng, H. Sun, X. Yan, H. Zhu, Z. Zou, S. Shen, and H. X. Liu, "Dense reinforcement learning for safety validation of autonomous vehicles," *Nature*, vol. 615, pp. 620–627, 2023.
[24] M. Bijelic, T. Gruber, F. Mannan, F. Kraus, W. Ritter, K. Dietmayer, and F. Heide, "Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020, pp. 11 679–11 689.
[25] Z. Zheng, Y. Cheng, Z. Xin, Z. Yu, and B. Zheng, "Robust perception under adverse conditions for autonomous driving based on data augmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 12, pp. 13 916–13 929, 2023.
[26] W. G. Najm, J. D. Smith, and M. Yanagisawa, "Pre-crash scenario typology for crash avoidance research," National Highway Traffic Safety Administration (NHTSA), Tech. Rep. DOT HS 810 767, April 2007, john A. Volpe National Transportation Systems Center, Cambridge, MA.
[27] Google DeepMind, "Gemini models," Technical Report, 2025, model used: `gemini-3-flash-preview`. Available: https://ai.google.dev/gemini-api/docs/models.
[28] X. Zhou and C. Alecsandru, "WayGraph: Intersection fingerprinting for GPS-free localization," 2026, in preparation.
[29] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proceedings of the Conference on Robot Learning (CoRL)*, 2017, pp. 1–16.
[30] Waymo LLC, "Waymo open dataset: End-to-end driving leaderboard," https://waymo.com/open/challenges/end-to-end-driving/, 2026, accessed: 2026-02-28.
[31] S. Shalev-Shwartz, S. Shammah, and A. Shashua, "On a formal model of safe and scalable self-driving cars," 2017.

Xingnan Zhou received the M.Eng. degree from the Chinese People's Public Security University, Beijing, China, in 2019. She is pursuing the Ph.D. degree at Concordia University, Montreal, Canada. Her research spans trajectory prediction, LiDAR–camera fusion, and VLM interpretability for autonomous driving. She is the developer of CTL-Drive (ranked 15th of 67 on WOD-E2E).

Ciprian Alecsandru received the Ph.D. degree from Louisiana State University in 2006. He is with the Department of Building, Civil and Environmental Engineering, Concordia University, Montreal, Canada, where he directs the Transportation Lab. His interests include traffic simulation, intelligent transportation systems, and autonomous vehicle safety.