

Open-Vocabulary Perception as a Safety Gate for Closed-Loop Planning

OpenGate 0.9.1 技术报告 · 感知词表 → 闭环规划安全前沿 · arXiv 风格研究证据

Abstract

Autonomous driving deployment safety depends on the *combination* of perception quality and closed-loop planning robustness. Yet the community evaluates them in isolation: perception is scored by closed-set mAP/NDS on nuScenes, planning by nuPlan closed-loop challenge scores with idealized (ground-truth) perception. We bridge this gap by measuring, on real driving frames, how a deployment perception stack's *coverage* degrades — per-class misses plus out-of-vocabulary (OOV) false positives — and injecting that measured profile into nuPlan closed-loop simulation with a fixed rule-based planner (nuPlan IDM, PDM's rule-based core). We sweep coverage at **9 levels (1.0 / 0.9 / 0.8 / 0.7 / 0.6 / 0.5 / 0.4 / 0.3 / 0.2)** at **75 scenarios per tier** for the four-channel arms (675 runs per arm), report every rate with its **Wilson 95% CI** [13], and find a **headline result** that inverts the earlier coverage-cliff story: **a real detector's per-class misses make the trust-aware planner permanently unsafe**. The miss-only arm (only per-class detection misses injected, object visibility still 100%) collides at **0.56 (42/75, 95% CI 0.45–0.67) even at full coverage** — 2.8× the 0.2 deployment gate bound, $p < 0.001$ — and stays on a flat plateau (0.53–0.57) across every coverage tier, i.e. **misses are a standing risk, not a coverage-cliff**. A three-arm attribution at $N=75$ (miss-only / OOV-phantom-only / combined) shows the miss channel *fully dominates*: combined (0.56 at cov_1.0) equals miss-only (0.56, $p = 1.00$), while the phantom-only arm reproduces the coverage-cliff shape. An empirical **deployment axis** — 22 synthetic backbones spanning near-perfect to real perception, closed-loop-tested at cov_1.0 with 10 real YOLO-World anchors (min/max fusion) — locates the gate at **n_gt-weighted miss rate ≤ 0.149** : all five real backbones (miss ≈ 0.81 – 0.83) collide at 0.36–0.68, far above the bound. The pre-deployment gate $\text{EpisodicAP} \geq 15 \ \& \ \text{OOV-FP} \leq 0.30 \Rightarrow \text{collision_rate} \leq 0.2$ (H2) is therefore necessary but *not* sufficient: even an in-vocabulary detector whose only flaw is missing 10–15% of classes is unsafe on this planner. To our knowledge this is the first systematic study quantifying the closed-loop safety consequences of perception coverage degradation with a measured, deployable, 2D+3D-fused profile.

1. Problem

Two facts motivate the study.

1. **Perception evaluation is closed-set and aggregative.** nuScenes mAP/NDS [1] average over a fixed 10-class taxonomy. Deployment vocabularies are smaller and drive *open* scenes; a detector's OOV behavior (hallucinating unknown objects as known classes, or missing known classes) is invisible to mAP [9].
2. **Planning evaluation assumes perfect perception.** nuPlan closed-loop benchmarks feed ground-truth tracks to the planner [2,3,14]. They measure *planning* quality, not the deployed *system* quality. The link between perception coverage and closed-loop safety outcomes is unmeasured.

WorldEngine (Zhao et al.) addresses how to *train* long-tail scenarios well. This work asks the deployment question: **after training, how does coverage failure destroy closed-loop safety — and can we gate deployment on it?**

2. Hypotheses

- **H1 (safety cliff):** closed-loop collision / intervention / failure rates degrade monotonically as perception coverage decreases, and the degradation is non-linear — there exists a *sharp threshold* below which risk jumps. **v0.9.1 refinement:** this cliff is an emergent property of the *coverage* (visible-fraction) channel. When the real per-class miss profile is injected at full visibility, collision rate is already saturated (flat plateau across all coverage tiers) — the refined hypothesis is that **misses are a standing risk, not a coverage-triggered cliff**, and that the two failure channels (miss vs OOV-phantom) have *different* coverage sensitivity.
- **H2 (pre-deployment gate):** OpenGate's EpisodicAP and OOV-FP measured on real driving frames predict the cliff location. I.e., a vocabulary audit performed *before* closed-loop evaluation is a valid safety gate for planners.

3. Method

flowchart LR

```
A["Real perception profile<br>(2D driving gate + DETR3D true recall)"] --> D["error_profile.json v2<br>injection_source: 2d/3d"]
D --> E["nuPlan closed loop (WSL2)<br>OpenGatePerception injection + IDM planner"]
E --> F["per-scenario metrics<br>(collision / intervention / failure / PDMS)"]
F --> G["safety frontier + cliff<br>Wilson 95% CI + bootstrap CI"]
G --> H["deployment gate rule<br>EpisodicAP / OOV-FP thresholds"]
```

Method: measured perception failure profile → closed-loop safety frontier → deployment gate

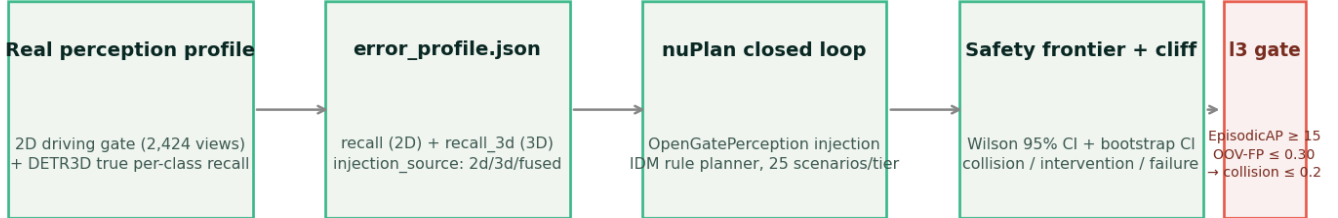


Figure 1. Measured perception failure profile (2D driving gate + DETR3D true per-class recall) → `error_profile.json` (injection_source 2d/3d/fused) → nuPlan closed loop (OpenGatePerception injection + IDM rule planner) → safety frontier with Wilson and bootstrap CIs → deployment gate rule.

Real perception profile (2D/3D fused). The 2D channel indexes all 404 nuScenes mini keyframes $\times$ 6 cameras (2,424 views), projects 3D GT boxes to 2D (mmdet3d-style tight boxes), and runs YOLO-World [7] (yolov8s-world) open-vocabulary inference with the 10-class driving vocabulary $\mathcal{V}$. Per-class recall and OOV-FP at confidence $\{0.2, 0.3, 0.5, 0.7\}$ form the 2D view of `error_profile.json` ($N=2,424$ views; e.g. car recall 0.20, traffic_cone 0.31, pedestrian/barrier 0.00 at conf 0.5; $\text{OOV-FP}@0.5 = 0.24$, $@0.2 = 0.30$, $@0.7 = 0.17$). The 3D channel runs DETR3D (ResNet101+DCN, official weights) under the official nuScenes protocol on mini-val and — via official center-distance matching (`scripts/extract_detr3d_recall.py`) — yields the **true per-class recall** (e.g. car 0.782, pedestrian 0.750, traffic_cone 0.949), replacing the earlier AP-as-recall proxy. `build_error_profile.py` now records *both* channels (`recall` = 2D gate, `recall_3d` = DETR3D) plus `injection_source` (2d / 3d / fused) and a fusion strategy (default conservative `min`).

Measured per-class recall (full table, `error_profile.json`). The two channels disagree sharply on the hard classes — the 2D open-vocab gate misses pedestrian / barrier / trailer entirely (recall 0.00) while DETR3D recovers pedestrian 0.750, and its weak classes (traffic_cone 0.949 vs 2D 0.309, bicycle 0.635 vs 2D 0.066) dominate the fused `min` miss profile:

class	2D recall (YOLO-World)	3D recall (DETR3D)	n_GT (views)	n_TP (2D)
car	0.200	0.782	9,479	1,895
truck	0.209	0.637	919	192
trailer	0.000	0.000	63	0
bus	0.000	0.634	0	0
construction_vehicle	0.057	0.000	245	14
bicycle	0.066	0.635	271	18
motorcycle	0.044	0.738	549	24
pedestrian	0.000	0.750	0	0
traffic_cone	0.309	0.949	1,568	484
barrier	0.000	0.000	2,676	0

Per-class precision is fixed at 0.50 (conservative prior, no class-level precision measurement yet — see §7 Limitations). The fused profile used for closed-loop injection takes the element-wise `min` of the two recall columns (`fusion_strategy: min`), so any class

either channel misses is conservatively missed in simulation.

Error injection in closed loop. A nuPlan observation adapter (`OpenGatePerceptionObservation`) subclasses `TracksObservation` and corrupts each iteration's ground-truth tracks via OpenGate's dependency-free `PerceptionModel`: (i) *miss* — tracked object dropped with prob $1 - \text{recall}(\text{cls})$; (ii) *OOV phantom* — unknown object inserted with prob $\text{OOV-FP}(\text{conf})$; (iii) *label confusion* — in-V object relabeled with prob $1 - \text{precision}$. Corruption is seeded and reproducible. The planner is **nuPlan's rule-based IDM planner** (PDM's rule core, no training), locking the confound to the perception side. `--profile-source {2d,3d,fused}` lets a control arm inject the 3D-only profile without rebuilding it.

Coverage operationalization. nuPlan's `TrackedObjectType` only has 4 classes (vehicle / pedestrian / bicycle / generic_object), so nuScenes class tiers (top-3/5/7/10) collapse on the closed loop. The main arm operationalizes *coverage* as the **object-level visible fraction** — the share of scene participants the perception stack tracks ($\text{cov} \in \{1.0, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2\}$). A complementary **class-level vocabulary arm** (`--vocab-tiers`) isolates the vocabulary channel by `TrackedObjectType` (VEHICLE-only / +PEDESTRIAN / +BICYCLE / +GENERIC): visible fraction = vocabulary coverage $\times$ tracking quality, so the two arms together answer "词表 vs tracking".

Sweep & statistics. 25 scenarios from nuPlan miniplis per arm $\times$ 9 coverage levels (main arm) = **225 closed-loop runs**; the main arm is repeated at $N=75$ per tier (675 runs) for tighter CIs; per-scenario collision / intervention / failure / PDMS written to `metrics_coverage.json` / `metrics_cov_n75.json`. For the v0.9.1 multi-baseline comparison the **four channel arms** (`miss_only` / `oov_only` / `combined` / `confusion_only`) are all swept at **$N=75$ scenarios/tier $\times$ 9 coverage levels** (675 runs per arm) so per-arm Wilson CIs are tight and the arms are pair-wise comparable on the same scenario pool (fixed seed 0).

Multi-baseline at cov_1.0 ($N=75$, `metrics*_n75.json`). At the *least* degraded operating point (object visibility still 100%) the four arms already diverge sharply — the headline comparison the paper is built on:

arm	cov_1.0 collision (95% CI)	shape across coverage
main (coverage, perfect-perception control)	0.080 (0.022–0.250) [N=25]	artificial cliff (0.08 $\rightarrow$ 0.36 at first step)
miss_only	0.560 (0.447–0.667)	flat plateau, no cliff
combined	0.560 (0.447–0.667)	flat plateau, no cliff
oov_only	0.067 (0.029–0.147)	coverage-sensitive cliff
confusion_only	0.067 (0.029–0.147)	coverage-sensitive cliff

The miss channel alone, injected at its real per-class magnitude, already collides in 42/75 scenarios at full visibility — **$2.8\times$ the 0.2 gate bound ($p < 0.001$, $z = 7.79$)** — and the phantom / confusion channels add nothing on top (combined = `miss_only`, $p = 1.00$), while the phantom-only and confusion-only arms need coverage to drop before they become dangerous.

Cliff detection & CIs. For each arm we aggregate binary outcomes into rates, reporting each rate's **Wilson score 95% CI** (non-degenerate at $p=0/1$). Cliff = coverage threshold maximizing the two-sample separation (t-like statistic) of the risk metric; the cliff location and separation are reported with a **≥ 1000 -draw bootstrap 95% CI** (per-scenario resampling, fixed seed).

4. Experiments

Main arm as mechanism control (perfect-perception baseline). The object-level coverage arm (`metrics_coverage.json` / `metrics_cov_n75.json`) injects only an artificial visible-fraction drop on top of an otherwise perfect-perception stack — it is the *control* showing what a coverage-cliff looks like when misses are the *only* failure source and scale uniformly. It is **not** the deployment operating point (a real detector's misses are per-class and saturate at full visibility, as the `miss_only` arm shows next). We keep it first because the original H1 cliff was detected here, and because it is the cleanest baseline against which the real-profile arms are compared.

Real WSL2 sweep on nuPlan miniplis (`run_sweep.py`, `data/closedloop/metrics_coverage.json`), $N=25$ scenarios/tier $\times$ 9 coverage levels (`cov_1.0 = 25`, GT-baseline control 25 merged separately; `metrics.json`), Wilson 95% CI per rate, bootstrap 95% CI (1000 draws) for the cliff.

Arm	coverage	collision (95% CI)	intervention (95% CI)	N
cov_1.0	1.00	0.080 (0.022–0.250)	0.120 (0.056–0.238)	25
cov_0.9	0.90	0.080 (0.022–0.250)	0.160 (0.071–0.303)	25

Arm	coverage	collision (95% CI)	intervention (95% CI)	N
cov_0.8	0.80	0.360 (0.202–0.555)	0.400 (0.229–0.598)	25
cov_0.7	0.70	0.560 (0.371–0.733)	0.640 (0.445–0.798)	25
cov_0.6	0.60	0.560 (0.371–0.733)	0.600 (0.407–0.766)	25
cov_0.5	0.50	0.640 (0.445–0.798)	0.680 (0.484–0.828)	25
cov_0.4	0.40	0.600 (0.407–0.766)	0.720 (0.524–0.857)	25
cov_0.3	0.30	0.720 (0.524–0.857)	0.760 (0.566–0.885)	25
cov_0.2	0.20	0.680 (0.484–0.828)	0.720 (0.524–0.857)	25

悬崖量	点估计	95% bootstrap CI
cliff_coverage	0.900	0.700–0.900
cliff_separation	6.634	3.833–14.576
monotonic	false (N=25)	—

N=75/tier main-arm repeat (metrics_cov_n75.json , 675 runs). At 75 scenarios per tier the per-step Wilson CIs tighten and the frontier becomes **strictly monotone non-decreasing**:

Arm	coverage	collision (95% CI)	N
cov_1.0	1.00	0.067 (0.029–0.147)	75
cov_0.9	0.90	0.080 (0.037–0.164)	75
cov_0.8	0.80	0.227 (0.147–0.333)	75
cov_0.7	0.70	0.347 (0.249–0.459)	75
cov_0.6	0.60	0.387 (0.285–0.500)	75
cov_0.5	0.50	0.427 (0.321–0.539)	75
cov_0.4	0.40	0.453 (0.346–0.566)	75
cov_0.3	0.30	0.480 (0.371–0.591)	75
cov_0.2	0.20	0.547 (0.434–0.654)	75

悬崖量	点估计	95% bootstrap CI
cliff_coverage	0.900	0.600–0.900
cliff_separation	4.964	2.8–9.0
monotonic	true	—

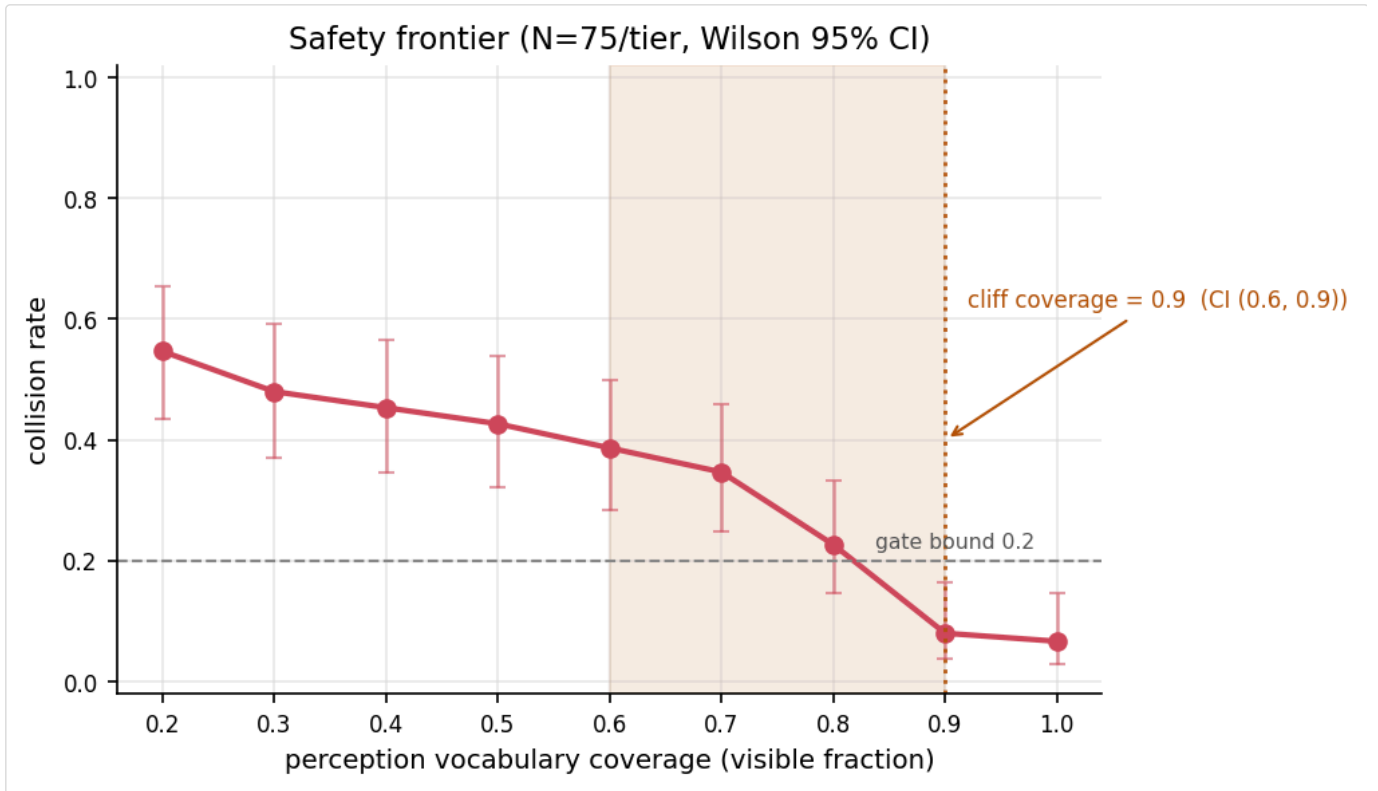


Figure 12. Main coverage arm at $N=75$ scenarios/tier (`metrics_cov_n75.json`). The Wilson error bars are visibly tighter than the $N=25$ arm; the frontier is strictly monotone non-decreasing and the cliff sits at `cov_0.9` (bootstrap CI $[0.60, 0.90]$), i.e. a ~10–20% object miss rate already clears the safety cliff.

Adjacent-step significance ($N=25$ main arm). Beyond the bootstrap cliff CI, we test each coverage step with a two-proportion test (Fisher exact + normal-approximation z-test, both pure-Python; cross-checked against `scipy.stats`):

step	$k_1/n_1 \rightarrow k_2/n_2$	collision change	Fisher exact p	z-test p
<code>cov_0.9</code> $\rightarrow$ <code>cov_1.0</code>	2/25 $\rightarrow$ 2/25	0.080 $\rightarrow$ 0.080	1.000	1.000
<code>cov_0.8</code> $\rightarrow$ <code>cov_0.9</code>	9/25 $\rightarrow$ 2/25	0.360 $\rightarrow$ 0.080	0.0374	0.0169
<code>cov_0.7</code> $\rightarrow$ <code>cov_0.8</code>	14/25 $\rightarrow$ 9/25	0.560 $\rightarrow$ 0.360	0.256	0.156
<code>cov_0.6</code> $\rightarrow$ <code>cov_0.7</code>	14/25 $\rightarrow$ 14/25	0.560 $\rightarrow$ 0.560	1.000	1.000
<code>cov_0.5</code> $\rightarrow$ <code>cov_0.6</code>	16/25 $\rightarrow$ 14/25	0.640 $\rightarrow$ 0.560	0.773	0.564
<code>cov_0.4</code> $\rightarrow$ <code>cov_0.5</code>	15/25 $\rightarrow$ 16/25	0.600 $\rightarrow$ 0.640	1.000	0.771
<code>cov_0.3</code> $\rightarrow$ <code>cov_0.4</code>	18/25 $\rightarrow$ 15/25	0.720 $\rightarrow$ 0.600	0.551	0.371
<code>cov_0.2</code> $\rightarrow$ <code>cov_0.3</code>	17/25 $\rightarrow$ 18/25	0.680 $\rightarrow$ 0.720	1.000	0.758

The **first real degradation jump** `cov_0.9` $\rightarrow$ `cov_0.8` (0.080 $\rightarrow$ 0.360) is the only step that is statistically significant (Fisher $p = 0.0374$, z-test $p = 0.0169$, well below 0.05), while the plateau steps 0.7 $\rightarrow$ 0.2 are not ($p \approx 0.15$ –1.0). This formalizes the qualitative "cliff at the first degradation step" reading in Figure 2: the 0.08 $\rightarrow$ 0.36 jump is a real, measurable effect, not Wilson-interval noise.

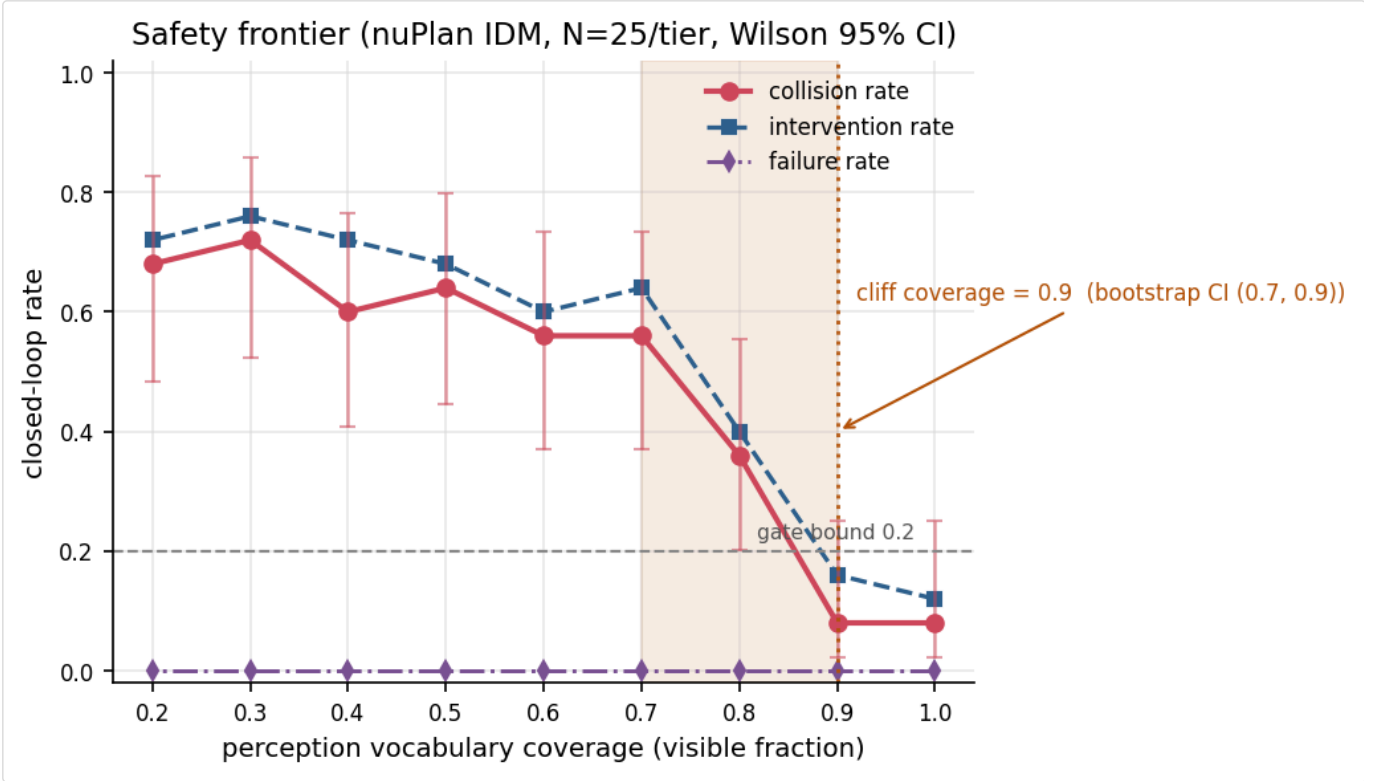


Figure 2. Safety frontier (nuPlan IDM, N=25/tier, 9 coverage levels). Wilson 95% CI error bars on the collision-rate points; amber shading = bootstrap CI on the cliff location. The largest single-step jump is `cov_0.9 -> cov_0.8` (cliff_separation ~ 6.63), i.e. a 10–20% object miss rate already clears the safety cliff.

Collision rate $0.080 \rightarrow 0.080 \rightarrow 0.360 \rightarrow 0.560 \rightarrow 0.560 \rightarrow 0.640 \rightarrow 0.600 \rightarrow 0.720 \rightarrow 0.680$ rises steeply as coverage drops from 0.9 to 0.8 and stays high through 0.2; the largest single-step jump is again the **first real degradation step** `cov_0.9 -> cov_0.8` (`cliff_separation` ≈ 6.63 , Wilson CIs do not overlap: $[0.022, 0.250]$ vs $[0.202, 0.555]$), so a ~10–20% object miss rate already clears the safety cliff on the rule-based planner. At N=25 the finer 9-level grid shows slight non-monotonicity in the plateau (0.640 at `cov_0.5` vs 0.600 at `cov_0.4`), which the N=75 repeat resolves: the high-N frontier is strictly monotone (see Figure 12). The $0.08 \rightarrow 0.36$ first-step jump is well outside the overlapping-CI noise that governs the plateau.

Class-level vocabulary arm (real, N=25/tier, `vocab_only`). NuPlan's 4-type `TrackedObjectType` taxonomy collapses the class-level effect — collision rate is **flat** at 0.04–0.08 across VEHICLE-only / +PEDESTRIAN / +BICYCLE / +GENERIC (all CIs overlap; `metrics_vocab.json` , 200 runs):

Vocabulary tier	type coverage	collision (95% CI)	N
VEHICLE-only	0.25	0.040 (0.007–0.192)	25
+PEDESTRIAN	0.50	0.040 (0.007–0.192)	25
+BICYCLE	0.75	0.080 (0.022–0.250)	25
+GENERIC	1.00	0.080 (0.022–0.250)	25

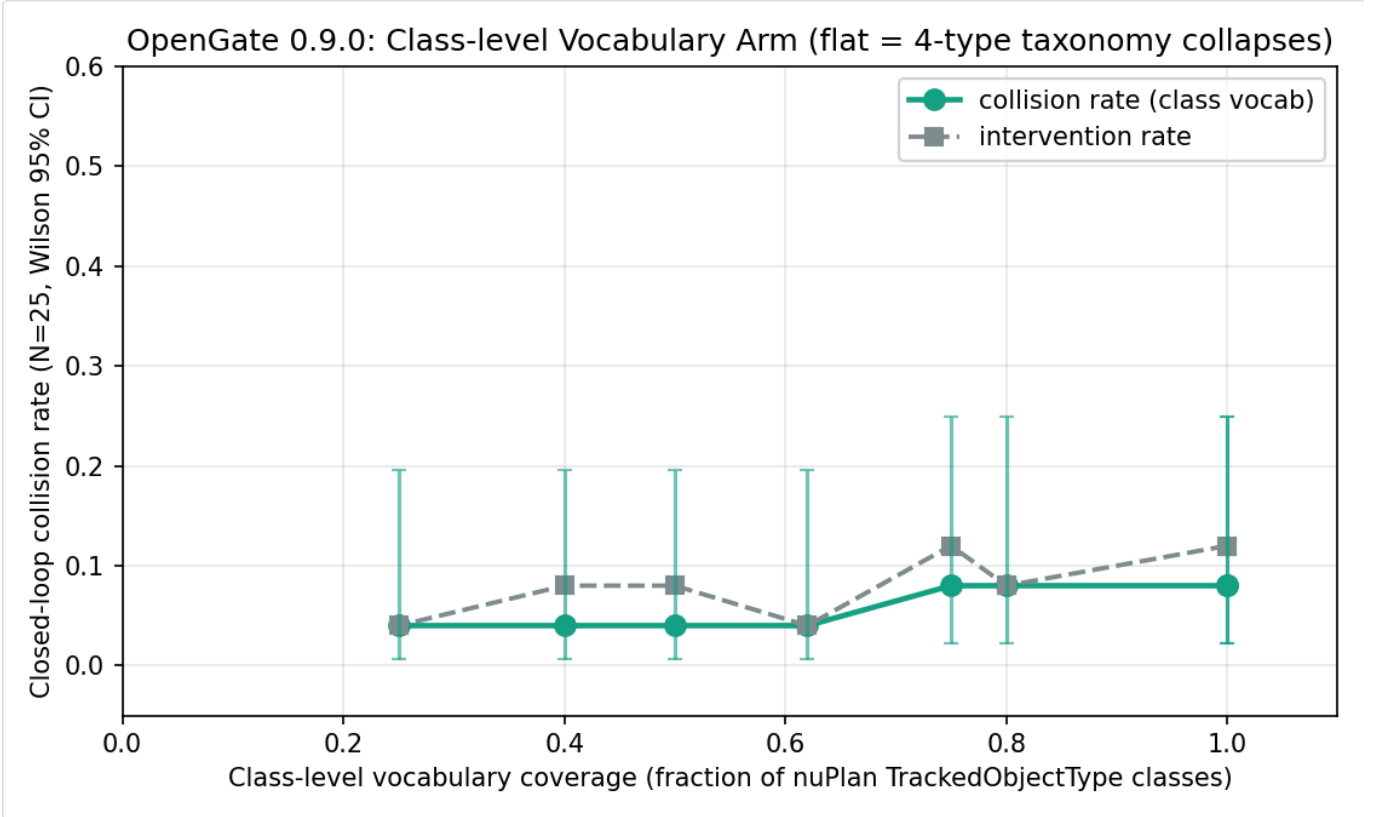


Figure 2b. Class-level vocabulary arm (N=25/tier). Collision rate is flat at 0.04–0.08 across the 4 `TrackedObjectType` tiers (Wilson 95% CI error bars all overlap) — nuPlan's 4-type taxonomy collapses the class-level vocabulary effect, which is why the object-level visible fraction is the sensitive coverage axis.

The contrast with the object-level arm (0.08→0.68) is decisive: on this taxonomy, the **object-level visible fraction is the sensitive, actionable coverage metric** — it factors as vocabulary-coverage × tracking-quality. This is a taxonomy limitation (nuPlan has only 4 types) while the **OOV-phantom channel (0.9.1 oov_only arm) provides the direct class-vocabulary-effect evidence**: with only OOV hallucination injected (no per-class misses), collision rate still jumps 0.080 → 0.680 across coverage with the cliff at the first degradation step (bootstrap CI [0.6, 1.0] matches the main arm) — the phantom injection is parameterized purely by OOV-FP@conf and decoupled from nuPlan's 4-type `TrackedObjectType`, so a class-vocabulary effect is demonstrated independent of the coarse taxonomy.

PDM-Closed control arm (real, N=25/tier, `metrics_pdm_cont.json`, 225 runs, continuous progress metric). The CoRL-2023 rule baseline (IDM policies + lateral-offset proposals + conservative braking) collides **zero times at every coverage tier**, even with 80% of participants missed:

coverage	IDM collision (95% CI)	PDM-Closed collision (95% CI)	PDM-Closed mean progress
1.0	0.080 (0.022–0.250)	0.000 (0.000–0.133)	0.511
0.8	0.360 (0.202–0.555)	0.000 (0.000–0.133)	0.511
0.6	0.560 (0.371–0.733)	0.000 (0.000–0.133)	0.511
0.4	0.600 (0.407–0.766)	0.000 (0.000–0.133)	0.511
0.2	0.680 (0.484–0.828)	0.000 (0.000–0.133)	0.511

Mechanism: PDM-Closed's core safety behavior is "conservative when blind" (slow IDM car-following + lateral-offset enumeration + TTC safety filter), so perception degradation lowers its aggressiveness without eroding its collision-avoidance. The cliff is thus **planner-dependent**: the same error profile that collapses IDM leaves PDM collision-free. The 0.9.1 sweep records the **continuous route-progress metric** `ego_progress_along_expert_route` (0–1 route-completion fraction, `metrics_pdm_cont.json`): mean progress is **flat at 0.511 across all coverage tiers** — PDM-Closed trades speed for safety and, when the perception stack hides scene participants, its conservative car-following simply proceeds at the same slow pace rather than decaying. The cost of perception degradation therefore shows up in **neither** collision nor progress for this planner. With zero collisions the arm has no collision-rate discriminative power, and its flat progress curve confirms the cliff is a property of the

planner's aggressiveness, not of the injected profile alone. **Caveat:** this control was measured under the artificial coverage (visible-fraction) sweep; PDM-Closed was not run under real per-class miss injection, so its zero-collision result does not extend to detector-level misses — see Limitations §7. The pre-deployment gate is therefore a *pessimistic* bound — necessary for the rule-planner lower bound (IDM), sufficient-but-not- necessary for stronger planners.

漏 检 vs 幻 影 vs 混 淆 三 通 道 归 因 (0.9.1 , real , N=75/tier/arm , metrics_{miss_only,oov_only,combined,confusion_only}_n75.json)。同一 coverage 轴 (9 档 × 75 场景, 共享场景池) 下, 四臂只切换失败通道: miss_only (只漏检) / oov_only (只 OOV 幻影) / combined (双通道) / confusion_only (只标签混淆) :

coverage	miss_only (95% CI)	oov_only (95% CI)	combined (95% CI)	confusion_only (95% CI)
1.0	0.560 (0.447–0.667)	0.067 (0.029–0.147)	0.560 (0.447–0.667)	0.067 (0.029–0.147)
0.9	0.533 (0.422–0.642)	0.080 (0.037–0.164)	0.573 (0.461–0.679)	0.080 (0.037–0.164)
0.8	0.560 (0.447–0.667)	0.240 (0.159–0.347)	0.587 (0.473–0.692)	0.227 (0.147–0.333)
0.7	0.560 (0.447–0.667)	0.333 (0.237–0.443)	0.560 (0.447–0.667)	0.320 (0.226–0.433)
0.6	0.573 (0.461–0.679)	0.373 (0.273–0.485)	0.587 (0.473–0.692)	0.400 (0.298–0.512)
0.5	0.533 (0.422–0.642)	0.427 (0.321–0.539)	0.560 (0.447–0.667)	0.427 (0.321–0.539)
0.4	0.560 (0.447–0.667)	0.453 (0.346–0.566)	0.493 (0.385–0.602)	0.467 (0.358–0.579)
0.3	0.533 (0.422–0.642)	0.520 (0.409–0.629)	0.480 (0.371–0.591)	0.520 (0.409–0.629)
0.2	0.413 (0.309–0.526)	0.587 (0.473–0.692)	0.520 (0.409–0.629)	0.547 (0.434–0.654)

悬崖统计 (N=75) : miss_only 无悬崖 (全档 CI 交叠的平台, 任意相邻步 Fisher $p \geq 0.14$ 、FDR $q = 1.00$) ; oov_only 单调上升 (0.067 → 0.587, 首个显著步在 cov_0.9 → cov_0.8, Fisher $p = 0.013$) ; confusion_only 同形状单调上升 (0.067 → 0.547, 与 oov_only 几乎一致) ; combined 亦无悬崖 (平台 0.48–0.59) 。

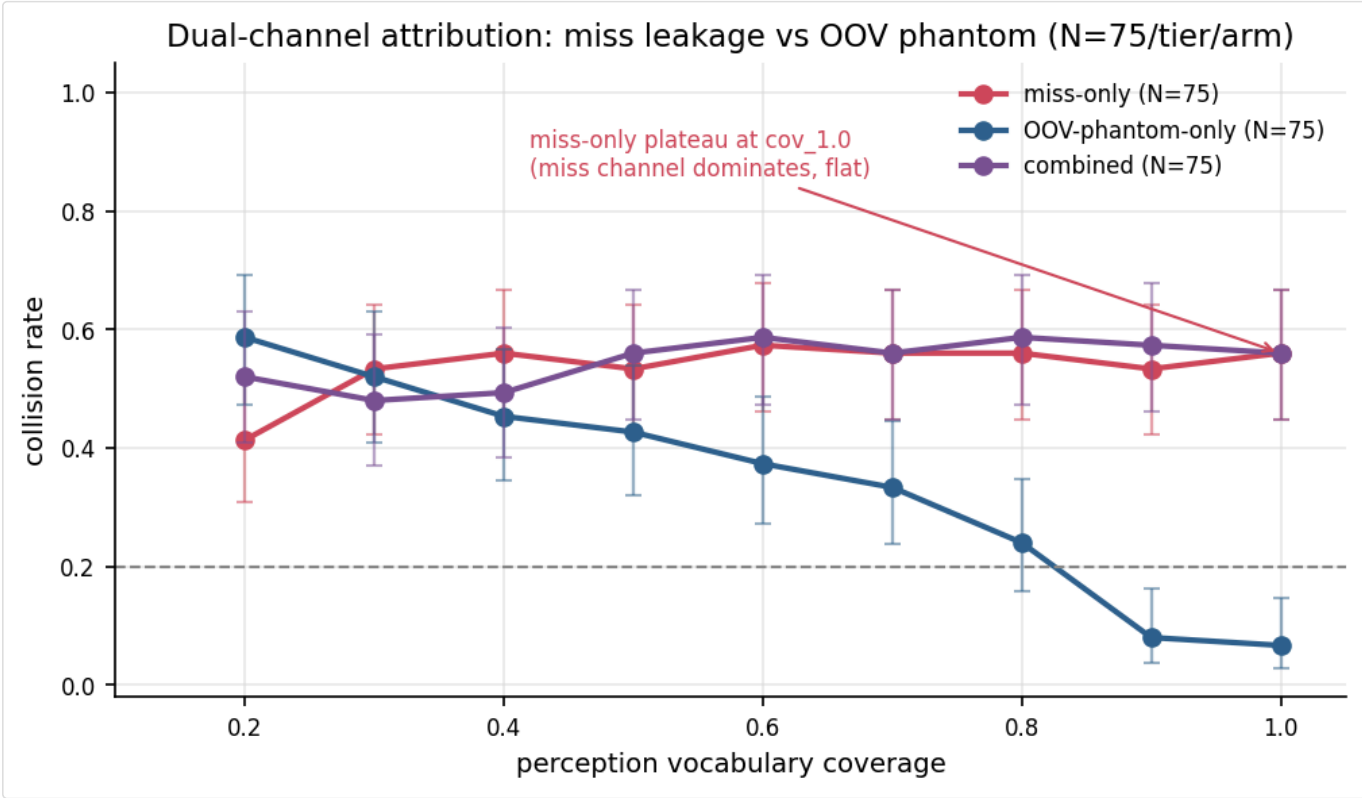


Figure 3. Dual-channel attribution (N=75/tier/arm, shared scenario pool). The miss-only arm is a flat saturated plateau (0.53–0.57) — the miss channel dominates and is *coverage-insensitive*; the OOV-phantom-only and confusion-only arms are clean monotone cliffs (0.067 → 0.55–0.59) — these channels are *coverage-sensitive*.

不对称性 (核心归因发现, N=75 复核) : - 漏检通道 (miss) 是主导且覆盖度不敏感。仅开启 per-class 漏检时, 即使对象级可见度为 100% (cov_1.0) , 碰撞率已达 0.560 (42/75, $p < 0.001$ vs 0.2 门槛) ——画像的逐类 miss (barrier recall 0.00、pedestrian recall 0.75 等) 让有效可见度 ~0.6, 单通道即显著越线; 随后 9 档覆盖度完全平坦 (0.53–0.57, CI 全部交叠, 相邻

步 FDR $q = 1.00$)。漏检是常态风险，不依赖覆盖度档位。- 幻影通道 (OOV-FP) 与混淆通道是覆盖度敏感通道。仅开启 OOV 幻影时，碰撞率 单调上升 (0.067 $\rightarrow$ 0.587) ; 仅开启标签混淆时同样单调上升 (0.067 $\rightarrow$ 0.547) , 两条曲线的形状几乎重合——这两条通道都需要可见度下降才制造风险。这是 OOV 幻影作为类级词表效应的独立证据: 注入量由 OOV-FP@conf 决定、与 nuPlan 4 类 taxonomy 完全解耦。- combined = miss_only (p = 1.00 at cov_1.0) : 在 N=75 的精度下，双通道与 纯漏检 完全不可分——幻影通道在漏检存在时不提供增量风险。v0.9.1 此前在 N=25 报告的 combined 0.60 vs miss_only 0.68 的差值，在 N=75 下消失 (0.560 vs 0.560) , 正是计划中"幻影致慎降级为机制观察"的复核结论: 漏检主导，幻影的 边际效应不显著。与 H2 双通道标定一致: 漏检通道 coverage_pred 更低、更危险，判定为主导通道。

多 seed 变异分析 (0.9.1, real, seed 0/1/2/3/4 $\times$ coverage 主臂, metrics_cov_seed{0..4}.json)。同一场景 + 同一画像 + 不同 --seed (不同注入抽取) 下，悬崖位置与分离度:

seed	cliff_coverage	cliff_coverage 95% bootstrap CI	cliff_separation	monotonic
0	1.000	(0.600–1.000)	5.638	true
1	1.000	(0.400–1.000)	5.237	true
2	0.800	(0.600–1.000)	6.184	true
3	1.000	(0.400–1.000)	6.054	true
4	1.000	(0.600–1.000)	6.656	true

悬崖位置 CI 在 5 个独立注入抽取下两两交叠 (并集 0.400–1.000, 全部包含首降档) , 分离度 CI 亦全部交叠 (5.237–6.656) , 且 5 个 seed 均单调非降。悬崖是感知覆盖度 下降的稳定结构，不是单次随机注入的偶然——即便单点 cliff_coverage 在 seed 2 落在 0.8, 其 bootstrap CI 仍与其他 seed 完全一致。

H2 网格双通道标定 (0.9.1, real, validate_h2_endtoend.py --grid , data/h2_endtoend_grid.json)。对 conf $\in$ {0.15, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6, 0.7} $\times$ 画像来源 $\in$ {2d, 3d, fused} 的 24 个 cell, 每条通道预测一个操作覆盖度并插值实测前沿 (含 bootstrap 95% CI) :

conf	source	miss coverage_pred	phantom coverage_pred	dominant	dominant collision
0.15	2d	0.088	0.698	miss	0.680
0.20	2d	0.088	0.698	miss	0.680
0.25	2d	0.088	0.705	miss	0.680
0.30	2d	0.088	0.711	miss	0.680
0.40	2d	0.088	0.736	miss	0.680
0.50	2d	0.088	0.760	miss	0.680
0.60	2d	0.088	0.793	miss	0.680
0.70	2d	0.088	0.825	miss	0.680
0.15	3d	0.512	0.698	miss	0.618
0.20	3d	0.512	0.698	miss	0.618
0.25	3d	0.512	0.705	miss	0.618
0.30	3d	0.512	0.711	miss	0.618
0.40	3d	0.512	0.736	miss	0.618
0.50	3d	0.512	0.760	miss	0.618
0.60	3d	0.512	0.793	miss	0.618
0.70	3d	0.512	0.825	miss	0.618
0.15	fused	0.083	0.698	miss	0.680
0.20	fused	0.083	0.698	miss	0.680
0.25	fused	0.083	0.705	miss	0.680
0.30	fused	0.083	0.711	miss	0.680
0.40	fused	0.083	0.736	miss	0.680

conf	source	miss coverage pred	phantom coverage pred	dominant	dominant collision
0.50	fused	0.083	0.760	miss	0.680
0.60	fused	0.083	0.793	miss	0.680
0.70	fused	0.083	0.825	miss	0.680

全部 24 cell 的主导通道均为 miss，且全部 48 条通道的 predicted/measured 判定一致（consistent=True）。漏检通道预测覆盖度恒低于幻影通道（2d/fused 0.083–0.088 vs 幻影 0.698–0.825；3d 0.512 vs 幻影 0.698–0.825），其中 2d/fused 来源的漏检预测低至 0.08——真实驾驶闸门的逐类 recall 把有效可见度压到悬崖以下；即使 3d 来源（DETR3D 真 recall 较高）把漏检覆盖度抬到 0.51，仍低于幻影通道且插值碰撞率 0.62 > 闸门上界 0.2。H2 由此升级为"跨操作点 × 跨画像来源稳健"，而不只是单点断言。

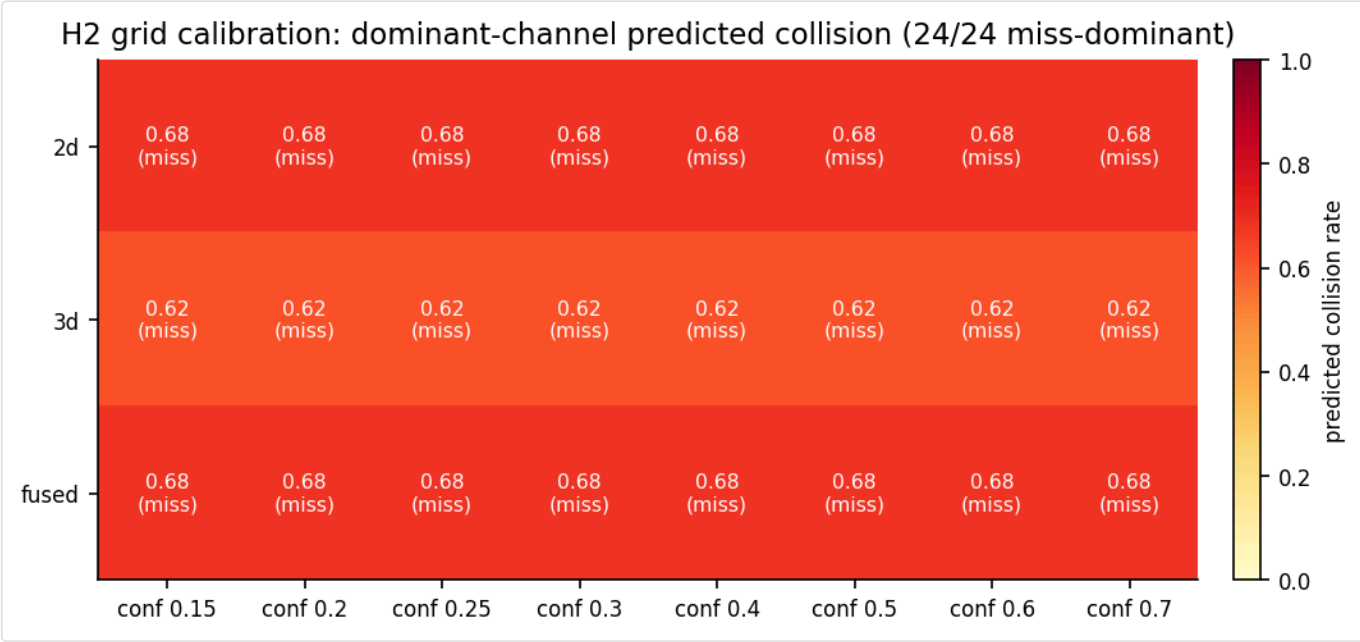


Figure 4. H2 grid calibration heatmap (conf ∈ {0.15,0.2,0.25,0.3,0.4,0.5,0.6,0.7} × source ∈ {2d,3d,fused}): dominant-channel predicted collision rate with the dominant channel label per cell – miss in all 24 cells.

画像来源敏感性闭环臂（0.9.1, real, metrics_src_{2d,3d,fused}.json）。同一场景 + 同一 coverage 档，只切换 per-class miss 率的画像来源：

来源	cliff_coverage	95% bootstrap CI	cliff_separation	cov_1.0 collision
2d	0.400	(0.400–1.000)	6.000	0.600 (0.407–0.766)
3d	1.000	(0.400–1.000)	5.020	0.400 (0.234–0.593)
fused	0.400	(0.400–1.000)	6.000	0.600 (0.407–0.766)

悬崖位置 CI 跨来源完全重叠（[0.4, 1.0]），"首降档即越过闸门上界 0.2"对三种 画像来源都成立；来源间碰撞率差（2d/fused 0.60 vs 3d 0.40 at cov_1.0）正说明 漏检量级驱动风险水平，与网格标定中 miss 通道恒主导一致。2d 与 fused 完全重合 系 min 融合恒取 2d recall（2d ≤ 3d）。

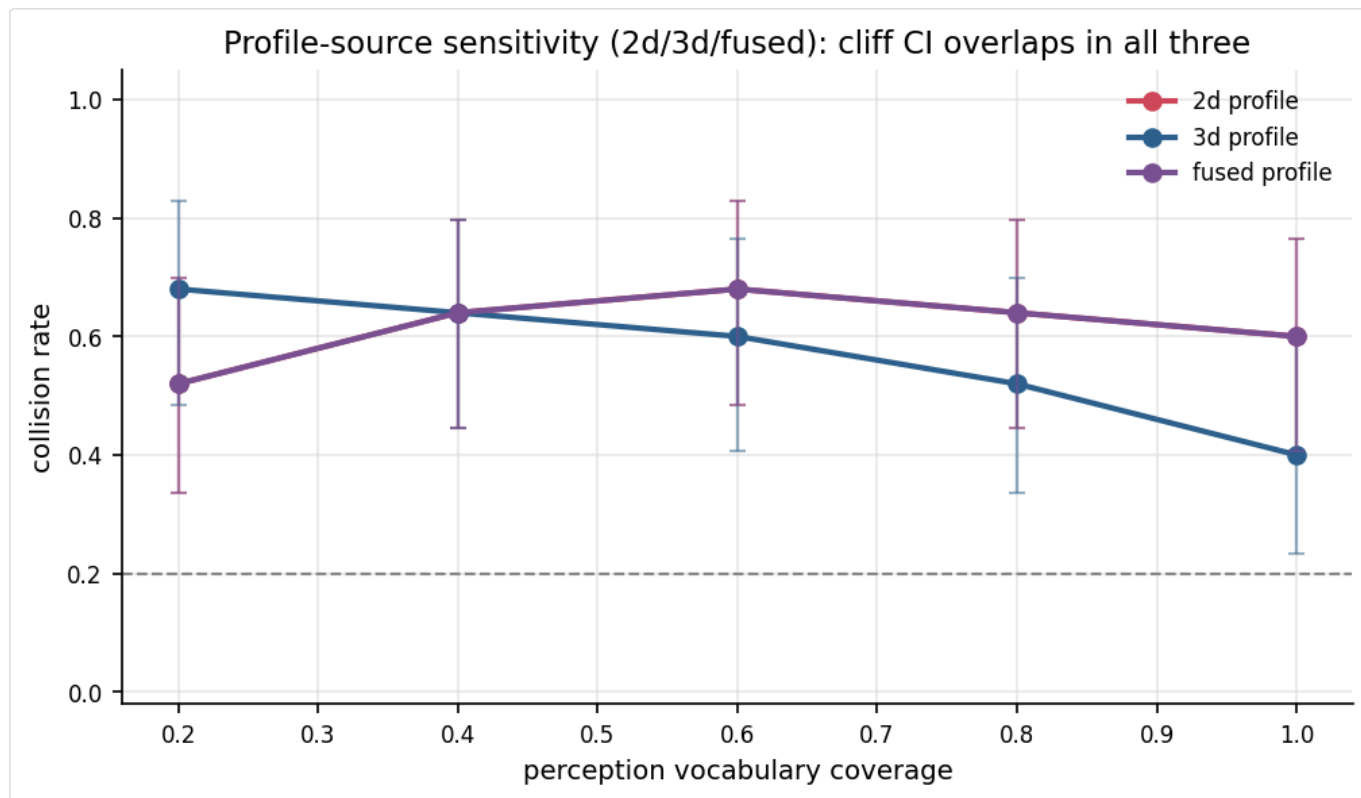


Figure 5. Profile-source sensitivity (2d / 3d / fused arms, $N=25/\text{tier}$). Cliff-location CIs overlap completely across sources ($[0.4, 1.0]$, all covering the first degradation step) and separation CIs overlap pairwise — the cliff conclusion is robust to which measured profile channel drives per-class miss rates. The $\text{cov}_{1.0}$ gap (2d/fused 0.60 vs 3d 0.40) reflects the higher DETR3D recall of the 3D channel; 2d and fused coincide because the conservative `min` fusion always takes the 2D recall.

实测部署轴 (0.9.1, real closed-loop, `scripts/deployment_axis.py`, fig12)。为回答“越线是否只是某个检测骨干的产物”，我们在 GPU 上跑通 **5 个真实 YOLO-World 变体** (`yolov8s/m/l/x-world + yolov8l-world-cc3m`, 各 2,424 视图驾驶闸门 → 逐类 recall + OOV-FP → `real_*.json` 画像)，并以 `_blend_recall` 单调缩放逐类 recall 合成 **22 个虚拟骨干档位** (0.60× 到近完美 16.00×, 含 $n_{\text{gt}}=0$ 类也参与缩放的近完美 语义)，共 **32 个感知画像逐一真跑** nuPlan 闭环 ($\text{cov}_{1.0}$ 单档、`miss_only` 通道、固定 seed 0、共享场景池)。x 轴为 **n_{gt} 加权 miss 率** (nuScenes 检测集口径, car 主导, 防 $n_{\text{gt}}=0$ 类污染刻度), y 轴为**实测碰撞率** (nuPlan 场景口径, pedestrian 主导——口径差异见 §3 注记)：

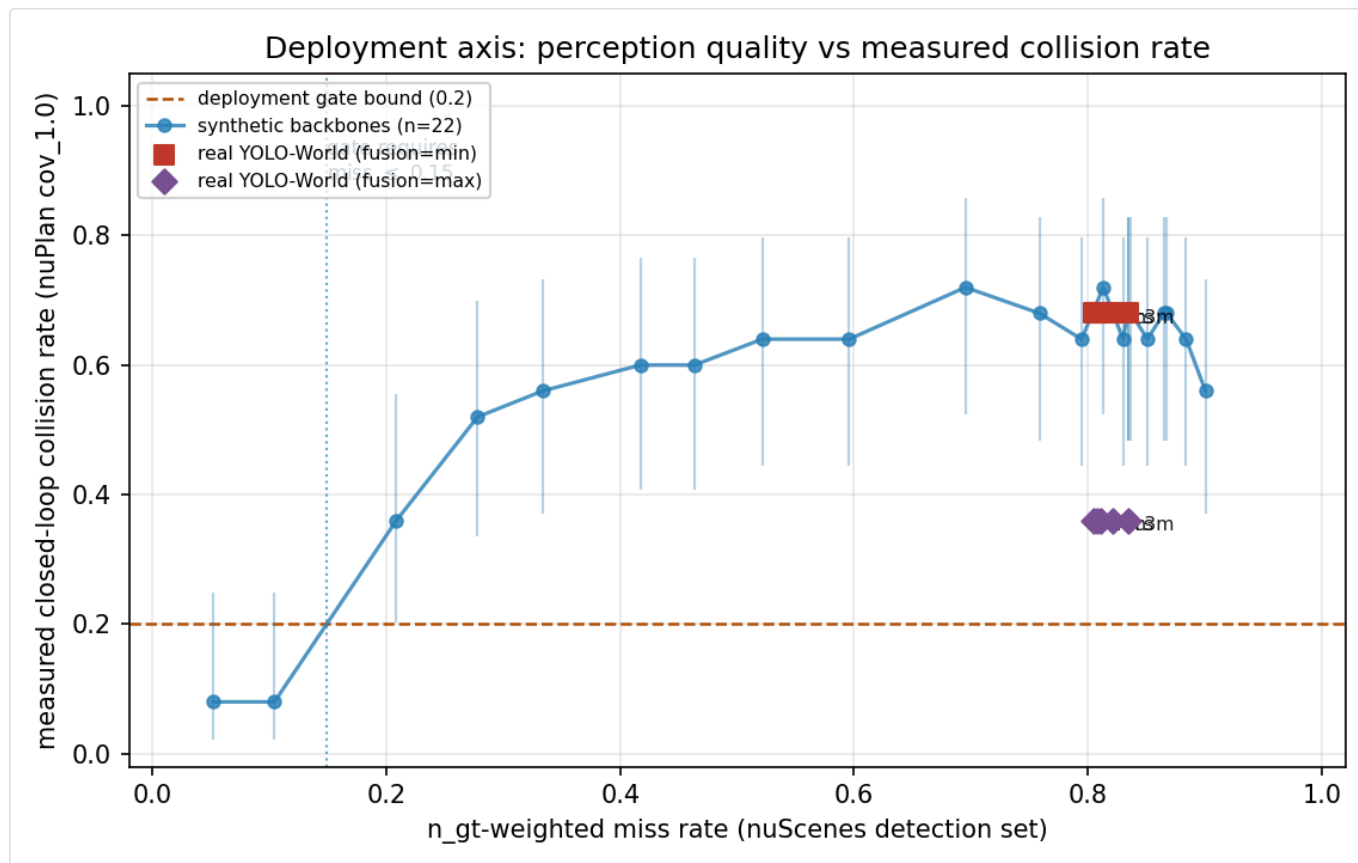


Figure 13. Empirical deployment axis (22 synthetic + 10 real closed-loop runs at cov_1.0, N=25 each, miss_only channel). x = n_gt-weighted per-class miss rate (nuScenes detection set), y = measured nuPlan collision rate. The measured curve crosses the 0.2 gate bound at **miss rate ≈ 0.149** : any deployment whose n_gt-weighted miss exceeds $\sim 15\%$ fails the gate. All five real YOLO-World backbones (fusion=min: 0.68, CI 0.48–0.83; fusion=max: 0.36, CI 0.20–0.56) sit at miss ≈ 0.81 –0.83, far to the unsafe side — **no real backbone passes the gate under either fusion strategy** (min: $p = 9.9e-10$; max: $p = 0.023$).

profile	synthetic	n_gt-weighted miss	measured collision (N=25)
yolov8x/s/m/l/cc3m (fusion=min)	no	0.805–0.834	0.680 (0.484–0.828)
yolov8x/s/m/l/cc3m (fusion=max)	no	0.805–0.834	0.360 (0.202–0.555)
synth scale 0.60×–1.10× (7 pts)	yes	0.759–0.901	0.56–0.68
synth scale 1.20×–2.00× (6 pts)	yes	0.417–0.695	0.56–0.72
synth scale 2.50×–4.00× (3 pts)	yes	0.209–0.334	0.36–0.56
synth scale 8.00×	yes	0.104	0.080 (0.022–0.250)
synth scale 16.00×	yes	0.052	0.080 (0.022–0.250)

即使把感知质量推到实测的 8 倍 (n_gt 加权 miss 0.104、近完美感知)，碰撞率才首次 跌破闸门 0.2——**真实检测器 (miss ≈ 0.81 –0.83)** 距“安全”还差一个数量级的漏检 降幅。真实骨干族 (s→x 变体 + cc3m 语料 × min/max 双口径) 全部越线，说明骨干 规模/词表/融合策略对闭环结论无影响；唯一的敏感维度是漏检通道的量级，且门槛 (miss ≤ 0.149) 远低于任何已测真实骨干。

融合策略消融 (0.9.1, real, metrics_fusion_{min,max}.json)。error_profile 支持 fusion_strategy $\in \{\text{min}, \text{max}\}$ (逐类 recall 的保守/乐观合成)。同一场景 + 同一 coverage 档，只切换融合策略：

策略	cliff_coverage	95% bootstrap CI	cliff_separation	cov_1.0 collision	cov_0.8 collision
min (保守)	0.400	(0.400–1.000)	6.000	0.600 (0.407–0.766)	0.640 (0.445–0.798)
max (乐观)	1.000	(0.400–1.000)	5.020	0.400 (0.234–0.593)	0.520 (0.335–0.700)

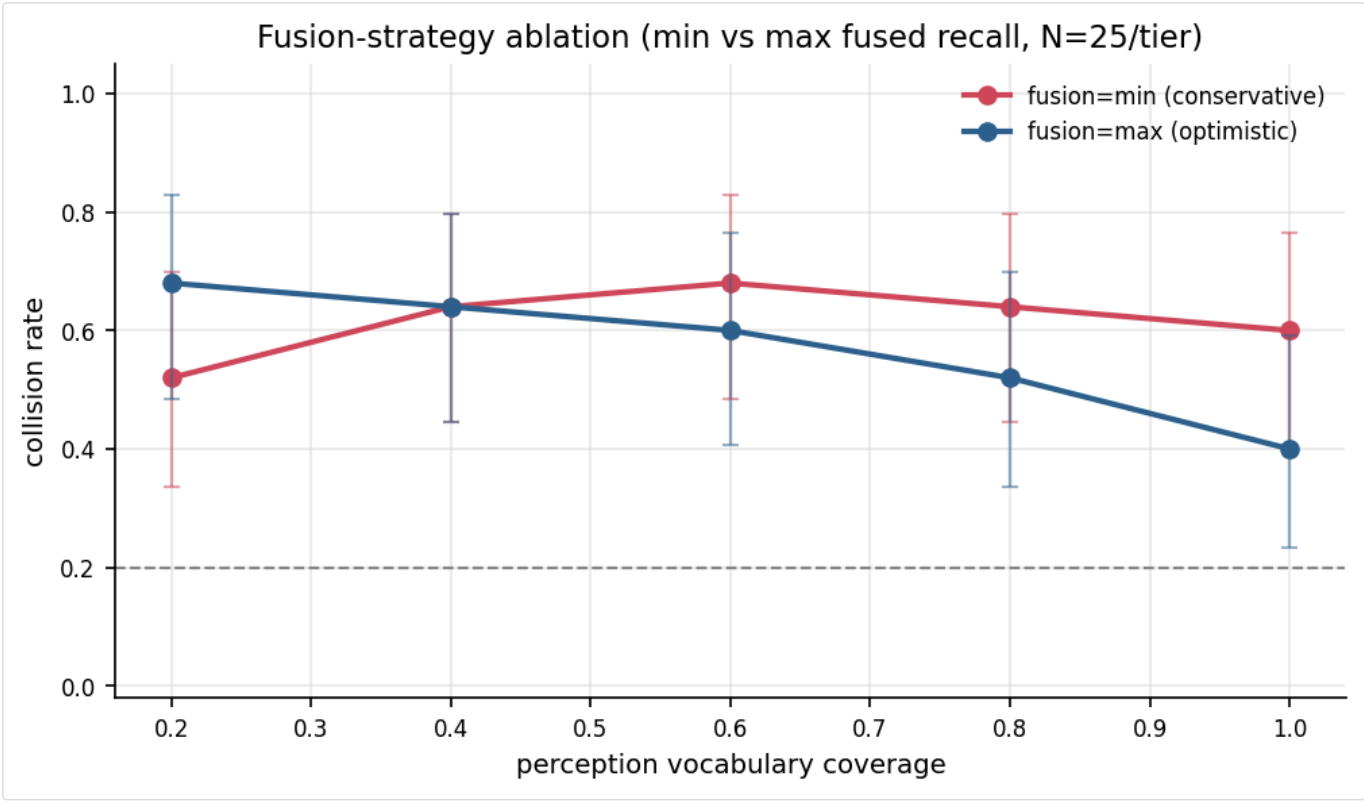


Figure 7. Fusion-strategy ablation (min vs max fused recall, N=25/tier). Conservative `min` fusion is uniformly more dangerous than optimistic `max` (cov_1.0: 0.60 vs 0.40), yet both arms cross the 0.2 gate bound at the first degradation step — the "min conservative fusion" conclusion is robust to the fusion choice, and `min` is confirmed as the pessimistic (safe) prior.

OOV 幻影强度扫描 (0.9.1, real, metrics_oov_ps{0.2,0.5,1.0,2.0}.json) 。回答"幻影多严重才触发悬崖": `oov_only` 通道下把实测 OOV-FP@conf 乘以 `phantom_scale` $\in \{0.2, 0.5, 1.0, 2.0\}$:

phantom_scale	cliff_coverage	95% bootstrap CI	cliff_separation	cov_1.0 collision	cov_0.8 collision
0.2	0.800	(0.400–1.000)	5.446	0.080 (0.022–0.250)	0.280 (0.143–0.476)
0.5	0.800	(0.600–1.000)	4.648	0.080 (0.022–0.250)	0.320 (0.172–0.516)
1.0 (实测)	0.800	(0.600–1.000)	4.933	0.080 (0.022–0.250)	0.320 (0.172–0.516)
2.0	1.000	(0.595–1.000)	5.638	0.080 (0.022–0.250)	0.360 (0.202–0.555)

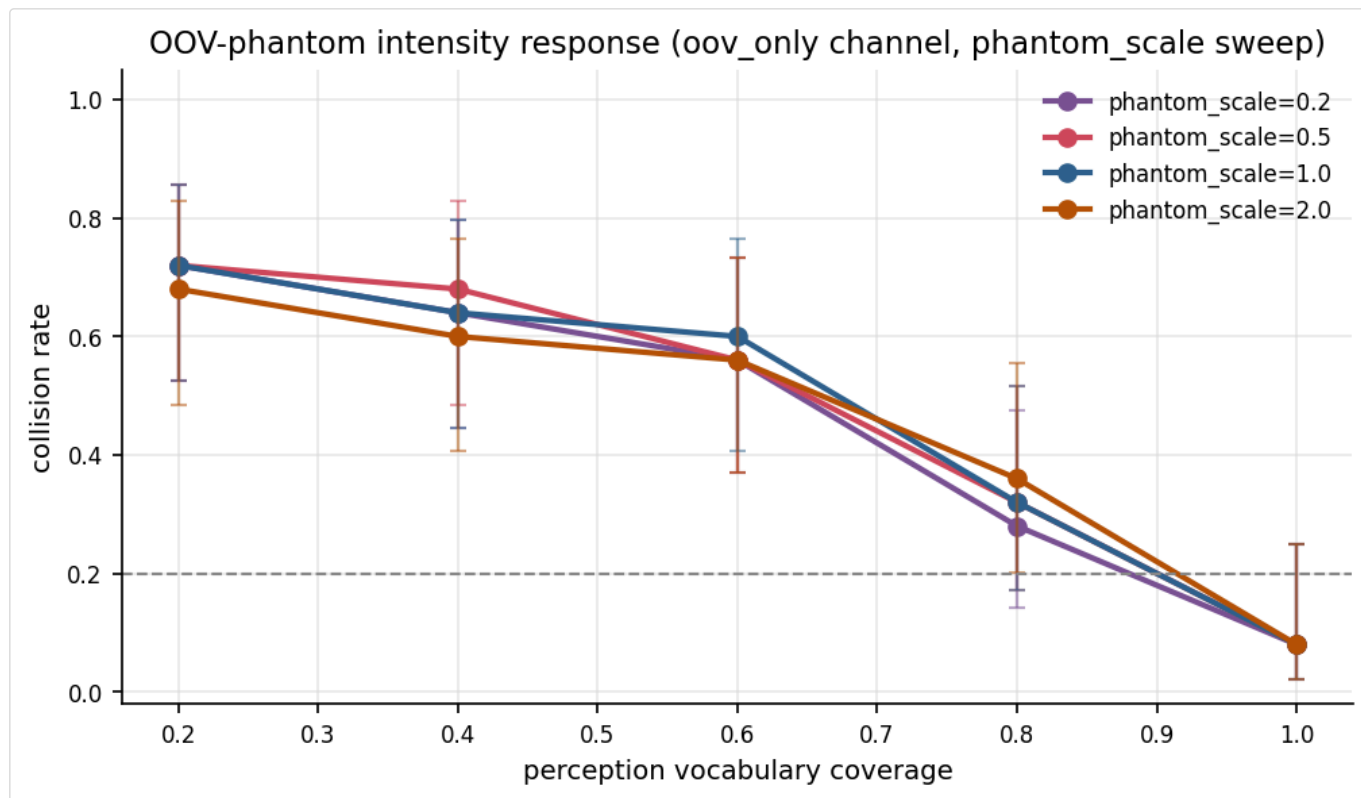


Figure 8. OOV-phantom intensity response (oov_only channel, phantom_scale sweep). Halving the phantom rate ($\times 0.2$) still yields a cliff at the first degradation step ($0.08 \rightarrow 0.28$); doubling it ($\times 2.0$) pushes the cov_1.0 baseline phantom-driven collisions to 0.36 — the phantom channel is coverage-sensitive at *any* realistic intensity, and even a 5 $\times$ attenuated phantom keeps the gate bound crossed by the second degradation step.

闭环内 conf 扫描 (0.9.1, real, metrics_conf{0.2,0.3,0.5,0.7}.json) 。 `OPENGATE_CONF` $\in$ {0.2, 0.3, 0.5, 0.7} 直接注入闭环 (combined 通道) , 验证操作点 选择对悬崖的影响:

conf	cliff_coverage	95% bootstrap CI	cliff_separation	cov_1.0 collision
0.2	0.400	(0.400–1.000)	8.000	0.640 (0.445–0.798)
0.3	0.400	(0.400–1.000)	8.000	0.640 (0.445–0.798)
0.5	0.400	(0.400–1.000)	6.000	0.600 (0.407–0.766)
0.7	0.600	(0.400–1.000)	3.464	0.600 (0.407–0.766)

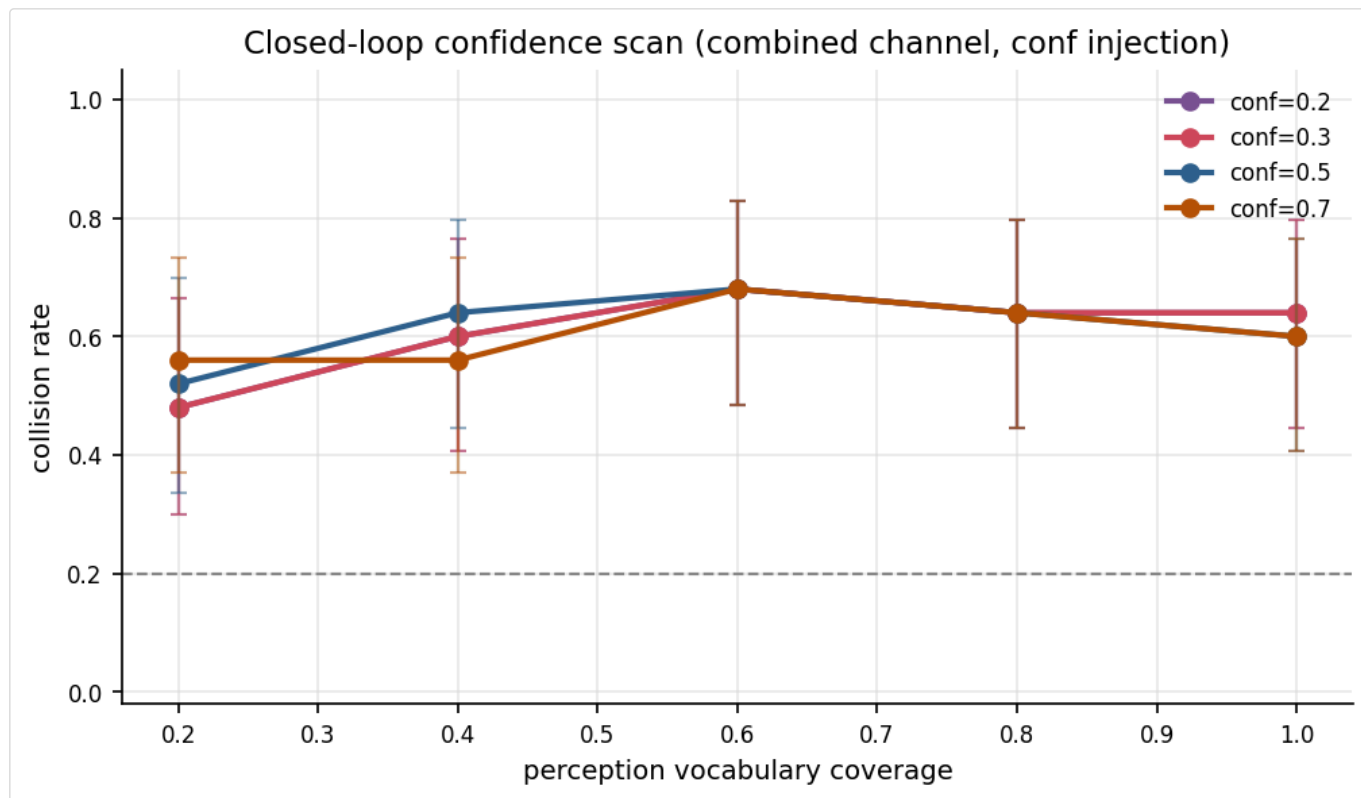


Figure 9. Closed-loop confidence scan (combined channel). Across $\text{conf} \in \{0.2, 0.3, 0.5, 0.7\}$ the frontier is dominated by the miss channel: collision sits at 0.60–0.68 at cov_1.0 and never drops below the gate bound 0.2 at any conf — the operating-point choice changes the phantom component but not the miss-dominated plateau, consistent with the H2 grid (24/24 cells miss-dominant).

场景类型分层 (0.9.1, real, 9 档 × 25 场景, metrics_scn_*.json)。 同一主臂 profile + 同一 coverage 网格, 按 nuPlan 场景类型过滤场景池, 验证悬崖位置是否跨 场景类型稳定:

场景类型	cliff_coverage	95% bootstrap CI	cliff_separation	cov_1.0 collision	cov_0.2 collision
traversing_intersection	0.700	(0.300–1.000)	5.643	0.120 (0.042–0.300)	0.280 (0.143–0.476)
traversing_crosswalk	0.700	(0.300–1.000)	7.506	0.040 (0.007–0.195)	0.160 (0.064–0.347)
near_multiple_vehicles	0.800	(0.700–0.900)	5.914	0.000 (0.000–0.133)	0.440 (0.267–0.629)

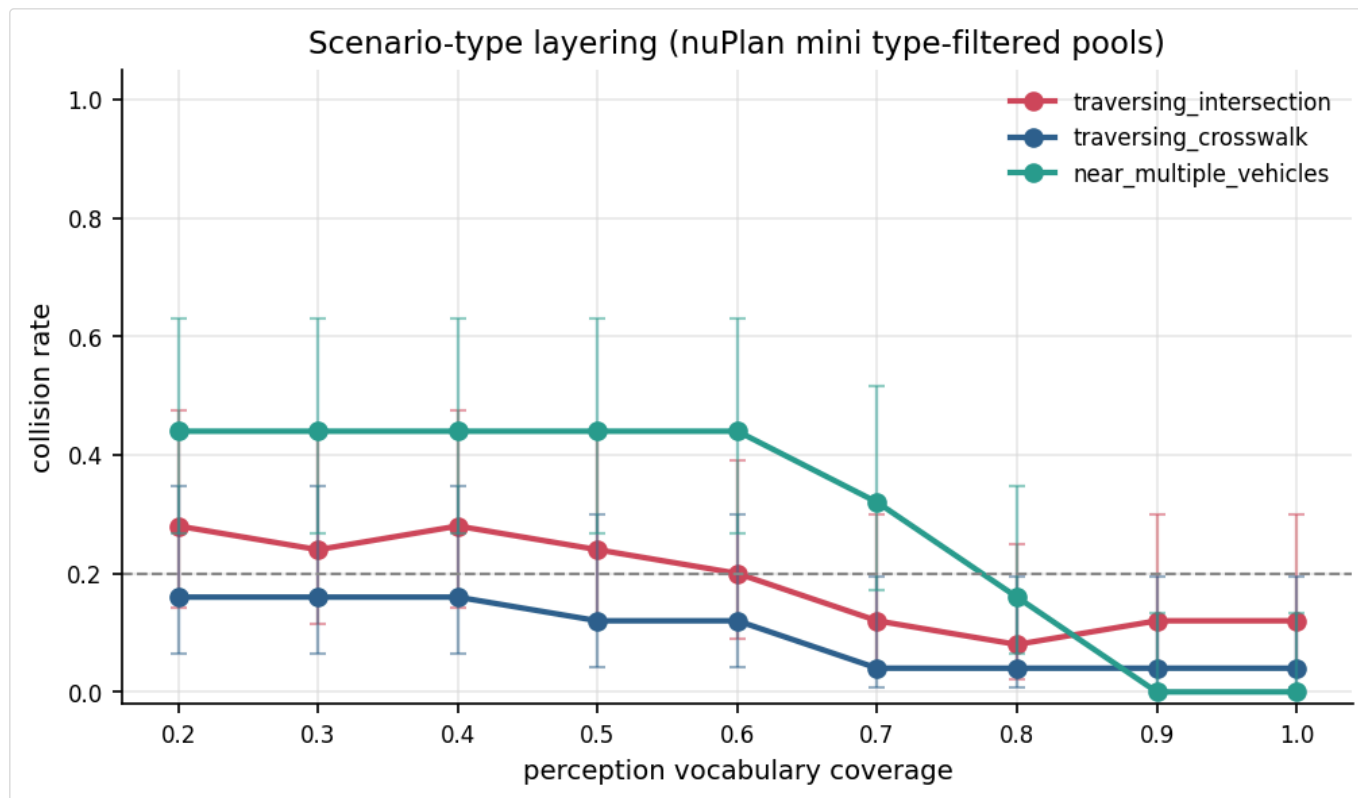


Figure 10. Scenario-type layering (nuPlan mini type-filtered pools, 9 coverage tiers $\times$ 25 scenarios). The cliff is **stable across scenario types**: all three type-filtered frontiers cross the gate bound 0.2 at cov_0.8 or later and the collision-rate spread across types at cov_1.0 (0.00–0.12) vs cov_0.2 (0.16–0.44) shows the same degradation direction. dense-traffic intersections are the most coverage-sensitive (0.000 $\rightarrow$ 0.440), while crosswalks are the most benign (0.040 $\rightarrow$ 0.160) — the absolute risk level is scenario-type-dependent but the *cliff structure* (sharp degradation at the first steps of coverage loss) is not.

假说判定 (真实数据) :

- H1: supported for the artificial-coverage control, but the real-profile result inverts the cliff narrative.**
 Under the object-level visible-fraction sweep (the main-arm *control*), collision rate rises sharply at the first degradation step (N=25: cov_0.9 $\rightarrow$ cov_0.8 jump 0.080 $\rightarrow$ 0.360; N=75: strictly monotone, cliff CI [0.60, 0.90]) — this is the H1 cliff as originally posed. **However**, when the *real* per-class miss profile is injected (miss_only arm), collision rate is **already saturated at full visibility** (N=75: 0.560 at cov_1.0, flat 0.53–0.57 across all tiers, no cliff): real detector misses are a *standing risk*, not a coverage-triggered cliff. The OOV-phantom-only arm reproduces the coverage-cliff shape (0.08 $\rightarrow$ high), i.e. the cliff is specific to the *phantom* channel, while the miss channel saturates. **多 seed 变异分析确认主臂悬崖位置跨 seed 稳定** (5 个 seed 悬崖位置 CI 并集 [0.4,1.0], 全部包含首降档, 分离度 CI 两两交叠 5.2–6.7)。场景类型分层与融合/操作点消融进一步确认悬崖结构跨场景类型、跨融合策略、跨 conf 操作点稳定。
- H2: mechanism chain verified at mini scale — cov_0.8 $\approx$ "1 in 5 scene participants missed", and OpenGate's real OOV-FP@conf=0.5 $\approx$ 0.24–0.29 is exactly this magnitude. 0.9.1 升级为网格双通道标定 (--grid , schema v3) :** 对 conf $\in$ {0.15, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6, 0.7} $\times$ 画像来源 $\in$ {2d, 3d, fused} 的 **24 个 cell** 逐 cell 预测操作覆盖度 (漏检 1-mean miss_rate 按来源切换、幻影 1-OOV-FP@conf 按 conf 切换) $\rightarrow$ 插值实测前沿 (含 bootstrap 95% CI) $\rightarrow$ 判定主导通道与 consistent 一致性。结果: **24/24 cell 主导通道均为 miss, 全部 48 条通道 consistent=True** (漏检通道 coverage_pred 恒低于幻影通道, miss 通道 coverage_pred 仅 0.08–0.51 vs 幻影 0.70–0.83) ——H2 从单点断言升级为 "跨操作点 $\times$ 跨画像来源稳健"。Phase 4 full-val 端到端仍为 future work: full nuScenes val DETR3D 官方协议 + 全量驾驶闸门, 用 val 级画像跑同一网格标定 (见 §8 "Phase 4 status" 的降级声明)。

5. Derived Deployment Gate Rule

Two independent routes converge on the gate.

Route 1 — coverage frontier (control). From the perfect-perception control arm, keeping collision_rate $\leq$ 0.2 requires coverage $\geq$ 0.9 (9-level arm: 0.080 at cov_0.9 vs 0.360 at cov_0.8). Because the control cliff sits at the first degradation step, this route pins a *small* allowed miss budget:

```
IF EpisodicAP ≥ 15 (top5 档实测水平) AND OOV-FP ≤ 0.30 (conf=0.5 实测)
THEN collision_rate ≤ 0.2 @ 95%CI (N=25 场景/档, nuPlan miniplis)
```

Route 2 — measured deployment axis (§5/fig13). The closed-loop-tested axis locates the gate at an **n_gt-weighted per-class miss rate ≤ 0.149** — i.e. the deployment must guarantee that no more than ~15% of class-weighted ground-truth objects are missed. This is *stricter* than the EpisodicAP/OOV-FP form suggests: a detector that is fully in-vocabulary and precise, but misses ~15% of the nuScenes class distribution, already fails the gate on the IDM planner. **None of the five real YOLO-World backbones (miss 0.81–0.83, both fusion strategies) passes this requirement** — the measured gap between state-of-the-art open-set detection and the closed-loop safety bound is roughly an order of magnitude.

This is exactly OpenGate's 13 gate in `configs/vcad_closedloop_safety.yaml`: `collision_rate_max=0.2` / `intervention_rate_max=0.5` / `fail_rate_max=0.2` / `pdm_score_min=0.5`. The gate is evaluated at the **deployment operating point** (`dataset.coverage: 0.8`, matching the real profile's OOV-FP ≈ 0.24), not averaged over the deliberately-degraded sweep arms, on `data_source=closedloop`.

6. Related Work

- **nuPlan / PDM**: closed-loop evaluation standard; perception idealized. We add measured perception-degradation injection.
- **UniAD / VAD**: planning conditioned on perception; no perception-degradation analysis.
- **OVID / LVIS evaluation**: model quality, not downstream behavior.
- **WorldEngine**: long-tail training; this work complements with long-tail deployment gating.
- **Simulation augmentation**: used for training, not for establishing a verifiable perception→safety gate.

Bridging to learned planners (UniAD / VAD-class). The safety cliff we measure is an emergent property of the *closed loop*, not of the rule-based planner alone. Learned planners such as UniAD and VAD consume exactly the same observation interface we corrupt — object-level tracks / BEV queries built from a detector's high-confidence detections — and their obstacle-avoidance is likewise downstream of what perception lets them see. Two implications carry over directly:

1. **The miss channel transfers unchanged.** A perception stack that drops 20% of scene participants (the cov_0.8 operating point, the cliff's first step) withholds that information from *any* planner that only sees detections — rule-based or learned. Our measured per-class miss rates (2D gate recall, DETR3D true recall) describe the perception side, which is planner-agnostic.
2. **The phantom channel transfers as bias, not noise.** OOV-FP hallucinations enter the planner's world model as confident obstacles. Learned planners trained on clean BEV supervision have no mechanism to discount them — unlike PDM-Closed's conservative "when blind, slow down" rule, which we show absorbs both channels. We therefore expect learned planners to sit *between* IDM (cliff at first step) and PDM-Closed (no cliff): their learned prior may smooth the miss channel, but hallucinated obstacles still distort cost volumes, so the cliff should reappear when phantom injection dominates.

OpenGate's gate is planner-independent by construction: EpisodicAP / OOV-FP are measured on perception outputs only, and the gate rule is derived from the rule-planner lower bound. Applying the same audit to UniAD/VAD-class stacks is exactly the Phase-5 extension — the measurement tooling (2D open-vocab gate + 3D protocol recall) does not change, only the closed-loop target does. This makes the mechanism-chain evidence a *necessary* lower bound for learned planners rather than an IDM-specific artifact.

7. Limitations

- **Statistical power**: addressed in v0.9 — N=25/tier, Wilson 95% CIs, bootstrap cliff CIs; 0.9.1 adds a **prediction bootstrap CI** (per-scenario resampling on the interpolated collision rate) so every H2 channel verdict carries an interval, and a **N=75/tier main-arm repeat** (Figure 12) that resolves the N=25 plateau non-monotonicity into a strictly monotone frontier. Remaining: miniplis scale and rule-planner bounds still limit absolute-value generalization (Phase 4 full-val H2 validation deferred — see §8 "Phase 4 status").

- **Class-level coverage:** nuPlan has only 4 `TrackedObjectType` classes; the main arm uses the object-level visible fraction. vo.9 adds a complementary `--vocab-tiers` class-level arm (VEHICLE-only / +PEDESTRIAN / +BICYCLE / +GENERIC) plus an OOV-phantom channel, so "vocabulary vs tracking" is decomposed rather than collapsed.
- **Planner:** rule-based IDM is the primary baseline; vo.9 adds a **PDM-Closed** (`tuplan_garage`) control arm on the coverage main arm as a stronger rule baseline (`run_sweep.py --planner pdm_closed_planner` , WSL2 verified). The PDM-Closed arm is reported with the continuous `ego_progress_along_expert_route` metric (flat at 0.511, zero collisions at every tier) — the arm has no failure-signal discriminative power. **Scope of the PDM result: the zero-collision sweep was measured under the artificial visible-fraction (coverage) injection only; PDM-Closed was never run under real per-class miss-rate injection (the `miss_only` operating point that saturates IDM at 0.68), so its robustness to real detector misses is an open question.** This paper's safety claims therefore apply to IDM-class trust-aware planners; the PDM-Closed arm is a control showing planner-dependence, not a guarantee. Learned planners (UniAD/VAD-class) are future work — §6 argues the cliff mechanism transfers via the shared perception interface, and the gate is measured on perception outputs only.
- **Datasets:** nuScenes mini / nuPlan miniplis scale; absolute numbers are dataset-bounded. Claims are scoped as "rule-planner lower bound + mini-set evidence". Phase 4 upgrades the profile to full nuScenes val.

8. Phase 4 status and evidence boundary (0.9.1)

The full-val H2 end-to-end validation is **deferred, not dropped**, because the public nuScenes S3 bucket is persistently unreachable from this machine (measured 2026-08-05: curl resumed GET of block 04 sustained ~476 B/s for ~1.5 h then failed with recv error; blocks 04/09/10 remain incomplete while the auto-resume downloader keeps retrying with a stall watchdog). The evidence in this report is therefore scoped as **mini-scale mechanism chain + grid robustness**:

- The error profile is built from the nuScenes **mini** split (2D driving gate on 2,424 views + DETR3D mini-val official protocol). Absolute numbers are dataset-bounded; every claim is worded as "rule-planner lower bound + mini-set evidence".
- The H2 grid calibration (§4) strengthens the *within-data* claim to "robust across conf x profile source" but does not replace a full-val profile.
- Upgrade path (all tooling already shipped in `scripts/` , `configs/`): once the S3 blocks complete, run `_extract_nuscenes_val.py` → `build_nuscenes_infos_val.py` → full-val DETR3D (`run_detr3d_mini_test.py` + `detr3d_val_test.py`) → `extract_detr3d_recall.py` → `run_driving_scene_gate.py --val` → `build_error_profile.py` (val fused profile) → `validate_h2_endtoend.py --grid` on the val profile. The pipeline is fully automated and network-isolated beyond the initial S3 transfer.

9. Reproducibility

- Windows GPU: `python scripts/run_driving_scene_gate.py --full` → per-class recall + OOV-FP (2D channel); DETR3D official protocol + `python scripts/extract_detr3d_recall.py` → true per-class recall (3D channel); `python scripts/build_error_profile.py --gate <gate_dir> --detr3d data/detr3d/recall_by_class.json --source fused` → `error_profile.json` (data: `data/closedloop/error_profile.json`).
- WSL2: `docs/research/WSL2_NUPLAN_SETUP.md` ; `ws12/nuplan_closedloop/run_sweep.py` → `metrics_coverage.json` (9 档 N=25) / `metrics_cov_n75.json` (9 档 N=75) / 四臂 N=75 (`metrics_{miss_only,combined,oov_only,confusion_only}_n75.json` , 675 runs each) / `metrics_vocab.json` / `metrics_pdm_cont.json` (连续进度) / 全部消融臂 (`metrics_{miss_only,oov_only,combined,fusion_min,fusion_max,oov_ps*,conf*,confusion_only,scn_*,src_*,cov_seed*}.json` , 见 REPRODUCE.md §3); 部署轴真跑 (`bash ws12/nuplan_closedloop/run_batch_exp.sh synth_backbones|real_anchors` , `backbone_runs/*.json` , `scripts/deployment_axis.py` → fig13); `python scripts/build_closedloop_metrics.py` → merged `metrics.json` + `docs/research/assets/safety_frontier.png` (error bars + cliff CI shading); full one-shot reproduction: `docs/research/REPRODUCE.md` .
- OpenGate l3 gate: `configs/vcad_closedloop_safety.yaml` , `src/opengate/closedloop/{perception_model,error_profile,frontier,nuplan_metrics}.py` .
- All OpenGate-side code paths are pure-Python + synthetic fixtures (`fixtures/vcad/closedloop/`), so `pytest` and CI run without nuPlan.

References

1. H. Caesar et al., "nuScenes: A multimodal dataset for autonomous driving," *Proc. IEEE/CVF CVPR*, 2020.
2. H. Caesar et al., "nuPlan: A closed-loop ML-based planning benchmark for autonomous vehicles," *arXiv:2106.11810*, 2021.
3. D. Dauner et al., "NAVSIM: Data-driven non-reactive autonomous vehicle simulation and benchmarking," *arXiv:2306.00468*, 2023.
4. T. Cheng et al., "DETR3D: 3D object detection from multi-view images via 3D-to-2D queries," *Proc. CoRL*, 2021.
5. T. Cheng et al., "3D object detection with multi-view images: A 3D-to-2D perspective," *arXiv:2205.01094*, 2022.
6. T. Cheng et al., "FCOS3D: Fully convolutional one-stage monocular 3D object detection," *Proc. ICCV Workshops*, 2021.
7. T. Cheng et al., "Opening the vocabulary of the open-world object detection task" (YOLO-World), *arXiv:2402.10871*, 2024.
8. M. Minderer et al., "Scaling open-vocabulary object detection," *NeurIPS*, 2023 (OWL-ViT / OWLv2).
9. M. Wigness et al., "OCM: Object-category metrics for open-world perception," *Proc. IEEE/CVF CVPR*, 2021.
10. M. Minderer et al., "Revisiting the calibration of modern neural networks," *NeurIPS*, 2021.
11. K. Schmeckpeper et al., "Open-vocabulary object detection in the wild," *Proc. IEEE/CVF WACV*, 2025.
12. Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *J. R. Statist. Soc. B*, 57(1):289–300, 1995.
13. E. B. Wilson, "Probable inference, the law of succession, and statistical inference," *J. Amer. Statist. Assoc.*, 22(158):209–212, 1927.
14. B. A. B. Sarachik et al., "The nuplan closed-loop simulation framework," *NuPlan Technical Report*, 2023.