FINDING AND CLOSING COVERAGE HOLES IN VISION-LANGUAGE MOTION PLANNERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language motion planners have advanced rapidly, yet single-trajectory accuracy can conceal missing feasible alternatives in imitation-trained planners. We define a *coverage hole* as a state–direction pair with verified trajectory witnesses, but no feasible sample from a reference planner at a fixed budget. Our verifier checks drivable-area compliance and collision-freeness against recorded agents, not traffic-rule compliance. We diagnose candidate holes using retimed cross-scene trajectory prototypes. The same witnesses supply augmentation targets, while a separate human-supervised detector selects the evaluation set. Neither training nor detection requires HD-map or maneuver annotations. Witness-target augmentation is combined with rehearsal and *prefix-scheduled decoding* (PSD), which allocates inference samples across direction prefixes and free generation using a scene-independent schedule. On nuScenes, we evaluate a 10B vision-language planner on 254 reference-defined hole pairs at 200 training-split states (about half of the pairs received a training target) and on validation and held-out scenes. Augmentation raises closure@128 from 46.5% to 85.0%, while minADE@8 decreases from 2.39 to 1.72 m. Adding PSD reaches **90.9%** closure [86.0,95.1] at 1.91 m minADE@8 and achieves 70% closure with about 29 samples; the off-the-shelf planner remains below 70% at 128. Validation closure under the official HD map reaches 96.0% versus 38.1%, with gains persisting under a lane-direction check. GT-only fine-tuning is more accurate (1.26 m) but recovers fewer reference-defined holes, and a matched-budget control with non-targeted witness targets closes most holes as well, so hole targeting mainly buys sampling efficiency at inference. These results show improved finite-budget candidate coverage under an open-loop verifier, with an accuracy trade-off relative to GT-only training.

1 INTRODUCTION

Vision-language motion planners generate future trajectories from camera observations and language context (Hwang et al., 2024; Wang et al., 2025b;c; Tian et al., 2024). These trajectories can serve as candidates for downstream selection, where a scorer or safety filter chooses among the available alternatives (Dauner et al., 2023; Jiang et al., 2026; Li et al., 2024). However, a selector cannot choose a maneuver absent from the candidate set, which makes candidate coverage important when no external route command specifies the intended direction. Consider an intersection where both turning left and turning right are feasible. A planner may generate only left turns, closely matching the driving log while omitting the feasible right turn. Imitation learning on driving logs supervises the recorded trajectory at each state without requiring the planner to represent these alternatives.

Generating diverse trajectories broadens the candidate set, but diverse candidates can still omit a specific direction, and aggregate diversity and admissibility metrics (Chen et al., 2024) do not answer this state-specific question. If a direction appears among the samples, it is evidently represented; the difficulty lies in the opposite case. When a direction is not generated, deciding whether this is a genuine omission requires evidence that the direction is feasible at that state, and the local driving log cannot provide it when it records only the executed maneuver.

We address this by using verified alternative trajectories as *witnesses*: constructive evidence that a direction is feasible at the current state. We define a **coverage hole** as a state–direction pair with such

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

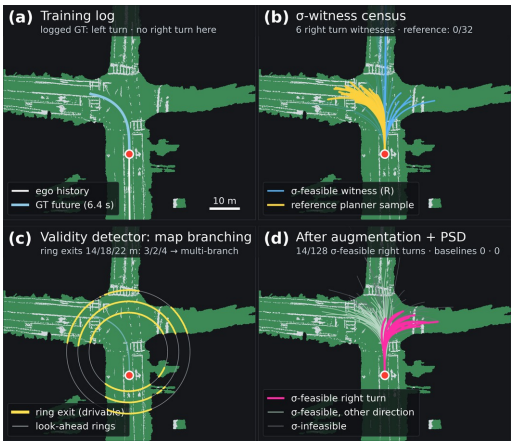


Figure 1: Diagnosing and recovering a coverage hole (missing right turn). **(a)** Training log. **(b)** σ-witness census: σ-feasible right-turn witnesses (blue) vs. reference-planner samples (yellow). **(c)** Validity detector: map-branching check (yellow = drivable ring exits). **(d)** After witness-target augmentation + PSD: σ-feasible right turns (magenta).

evidence but no feasible sample in that direction within a fixed sampling budget from a reference planner. Our feasibility predicate σ checks drivable-area compliance and collision-freeness against recorded, non-reactive agents; it does not establish traffic-rule compliance or interactive safety. A hole therefore denotes an observed omission under this verifier and budget, not absence from the model's support. Our central idea is to use the same witnesses both to diagnose these omissions and to supervise recovery of the missing directions.

Following MultiPath (Chai et al., 2019), we cluster training trajectories into representative anchors, retime them to the current state, and verify them under σ using a camera-derived map and recorded agent tracks. Comparing these witnesses with GT-finetuned Alpamayo samples identifies candidate holes (Fig. 1). A human-supervised validity detector selects evaluation cases. We propose to recover missing directions by fine-tuning off-the-shelf Alpamayo on witness targets paired with direction sentences and mixed with original-data rehearsal. Neither witness-target training nor hole detection requires HD-map or maneuver annotations. Prefix-scheduled decoding (PSD) further improves recovery by allocating samples across direction prefixes and free generation using a fixed, scene-independent schedule without an external route command.

On nuScenes, we measure recovery by closure@128: the fraction of fixed reference-defined hole pairs recovered by at least one of 128 independently drawn samples satisfying σ. Off-the-shelf Alpamayo closes only 46.5% of these pairs on the training-split evaluation set and 38.1% on validation scenes scored with the official HD map. Our method raises these rates to 90.9% and 96.0%, respectively. Coverage gains persist under a lane-direction check and on held-out scenes. A matched-budget non-targeted witness control recovers most holes as well, showing that hole targeting primarily improves sampling efficiency at inference. These gains involve trade-offs: augmentation is less accurate than GT-only fine-tuning, and PSD further reduces accuracy and sample feasibility. These results concern candidate coverage under the stated open-loop verifier, not rule compliance across the generated sample set.

Contributions.

- We formulate coverage holes as state–direction omissions relative to a reference planner and a finite sampling budget, and identify them using verified cross-scene witnesses.
- We use the same witnesses as augmentation targets and introduce PSD, a scene-independent prefix schedule, for finite-budget recovery.
- Through controlled comparisons, we distinguish the recovery gains from witness augmentation and PSD from the inference sampling-efficiency gains of hole targeting, and quantify their accuracy and lane-direction trade-offs.

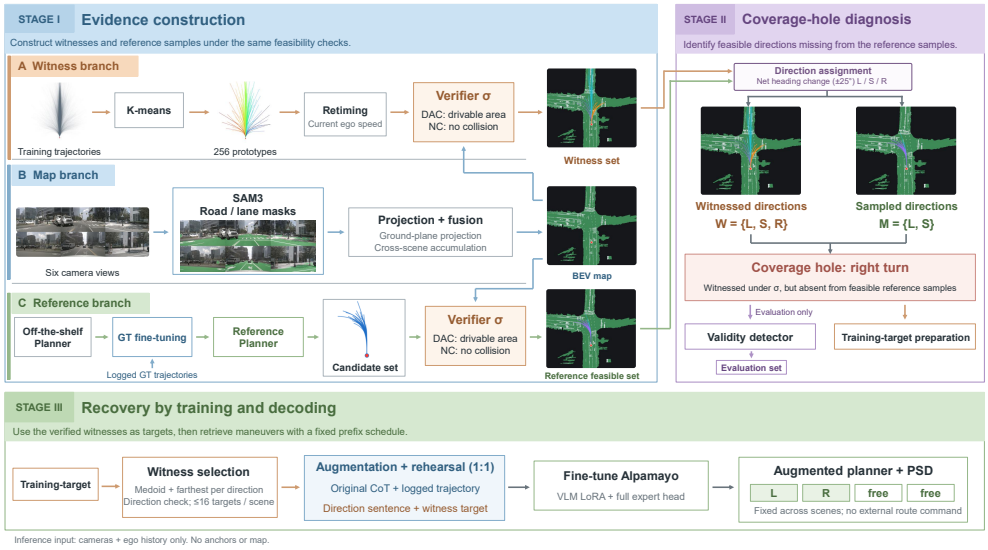


Figure 2: Overview. **Top:** retimed prototypes and reference-planner samples are verified under σ and compared by maneuver direction; here six witnesses establish right-turn feasibility, but none of the 32 reference samples realizes a σ-feasible right turn. The validity detector selects evaluation cases; training targets come from the full candidate pool. **Bottom:** witness-target augmentation with 1:1 rehearsal, fine-tuning and prefix-scheduled decoding. Inference uses cameras and ego history only; σ checks geometry and recorded-agent clearance, not traffic rules.

2 Related Work

Vision-language motion planning. EMMA (Hwang et al., 2024), OmniDrive (Wang et al., 2025b), DriveVLM (Tian et al., 2024), and Alpamayo (Wang et al., 2025c) incorporate language into driving scene understanding and motion planning. Grounded reasoning can improve control (Zawalski et al., 2024), but language-level reasoning does not guarantee consistency with the resulting action (Mayumu et al., 2026); counterfactual relabeling and consistency-aware planning address related alignment problems (Glossop et al., 2026; Yang et al., 2026). We instead ask whether feasible maneuver directions are represented within a fixed sampling budget, and combine supervision for missing directions with prefix-scheduled decoding.

Coverage in multimodal trajectory generation. Mode averaging and collapse can limit the alternatives represented by trajectory generators (Makansi et al., 2019; Hugenholtz et al., 2025). Anchors (Chai et al., 2019) and inverse-frequency rebalancing (Wang et al., 2025a) target underrepresented behaviors, and diversity and admissibility metrics go beyond trajectory error (Chen et al., 2024); yet aggregate metrics can obscure failures (Liu et al., 2025a; Sun et al., 2026) and do not show whether a specific feasible direction is represented at a given state. We define coverage holes using constructive feasibility evidence and finite-budget samples from a reference planner. Our closure metric is related to generative precision–recall (Kynkäänniemi et al., 2019), but measures recovery over these fixed holes rather than recall over the complete feasible set.

Candidate generation and selection. PDM (Dauner et al., 2023), VADv2 (Jiang et al., 2026) and Hydra-MDP (Li et al., 2024) generate and score candidates, and planning-aware diversity objectives seek alternatives useful downstream (Cui et al., 2021). Inference scaling helps (Baniodeh et al., 2025), but finite-budget success stays limited when useful candidates have low probability (Schaeffer et al., 2025; Brown et al., 2024), and a verifier (Kwok et al., 2025) cannot recover an absent maneuver. CLOVER (Ang et al., 2026) trains a proposal generator with evaluator-filtered pseudo-expert coverage supervision and a learned scorer; we instead diagnose reference-relative omissions at a fixed sampling budget and separate the effects of witness supervision, prefix scheduling and hole targeting.

Trajectory augmentation and guided generation. Counterfactual relabeling improves instruction following in VLAs (Glossop et al., 2026), while counterfactual scenario generation and reinforcement fine-tuning support driving evaluation and policy adaptation (Jing et al., 2026; Chen et al., 2026); condition-dependent source distributions (Dong et al., 2025), building on general flow and interpolant formulations (Tong et al., 2023; Albergo et al., 2025), also shape flow-based generation. Constraint-guided samplers (Liu et al., 2026; 2025b) and language-guided diffusion (Ding et al., 2026) shape generation at inference, while DIVER (Song et al., 2026) uses reward-driven training to improve trajectory diversity and driving quality. Our approach constructs augmentation targets offline from verified cross-scene trajectories and uses a fixed prefix schedule at inference, leaving the trajectory sampler and its flow-step budget unchanged. We separately evaluate witness augmentation, decoding, and hole targeting.

3 METHOD

3.1 PROBLEM DEFINITION

We study whether a vision-language planner represents feasible maneuver directions within a finite sampling budget; Fig. 2 gives an overview. A *state* x is a scene–timestamp pair with observation o_x, consisting of camera views and ego history. A planner $p_\theta(\tau \mid o_x)$ generates a future trajectory τ from this observation without an external route command. Each non-stop trajectory is assigned a maneuver direction $D(\tau) \in \{\mathrm{L}, \mathrm{R}, \mathrm{S}\}$ from its net heading change (thresholds and minimum length: Appendix A.1.1).

A trajectory is σ**-feasible** if it satisfies drivable-area compliance on the map used for verification and no-collision against recorded agent tracks; the map and tracks of state x are implicit in σ. The predicate checks neither lane direction nor other traffic rules, and the recorded agents do not react to the generated trajectory, so it does not assess interactive safety.

Let $\mathcal{W}_d(x)$ denote the set of *witnesses* for direction d at state x: σ-feasible trajectories constructed from retimed cross-scene prototypes, rather than samples from the reference planner (§3.3). A direction is witnessed if $|\mathcal{W}_d(x)| \geq 1$. We use only states with at least two witnessed directions, denoted $\mathcal{X}_{\mathrm{multi}}$, so that several alternatives are available at each state. At each state we draw N_{det} free-generation samples $\tau_{1:N_{\mathrm{det}}}$ from a fixed *reference planner*. A witnessed direction forms a candidate coverage hole if none of these samples realizes it while satisfying σ:

$$\mathcal{H}_{\mathrm{ref}} = \Big\{(x, d) :\ x \in \mathcal{X}_{\mathrm{multi}},\ |\mathcal{W}_d(x)| \geq 1,\ \textstyle\sum_{i=1}^{N_{\mathrm{det}}} \mathbf{1}[\sigma(\tau_i) = 1 \wedge D(\tau_i) = d] = 0\Big\}. \quad (1)$$

A coverage hole depends on the witnesses, verifier, reference planner, decoding protocol and finite detection run: it is an observed omission within that budget, not evidence that the direction lies outside the model's support.

Our goal is to recover these missing directions within a separate generation budget N: for a fixed set of hole pairs, recovery succeeds when at least one independently drawn trajectory realizes the missing direction while satisfying σ (§4). We ask how witness-target training and inference-time decoding affect this recovery.

3.2 SENSOR-DERIVED MAP

We construct the BEV map in three stages. First, we obtain road/lane masks from the six camera views using SAM3 (Carion et al., 2026). We back-project these masks onto the ground plane and accumulate them across each scene to produce a per-scene occupancy map (v1). Second, we fuse the per-scene maps in world coordinates using all training-split scenes that observe the same location (v2). Locations seen by at least two scenes are added if at least 40% of them mark it drivable; locations seen by at least four scenes are removed if drivable support is below 15%. The fused map v2 is used for all training-side σ evaluations; a smoothed variant with lane fusion (v3) is used for visualization and for the alternative witness construction of Appendix A.3.6.

These maps accumulate full-scene logs and serve only offline supervision and evaluation, never planner input; validation scenes have no fused map and use the official HD map; Fig. 3 shows the map versions.

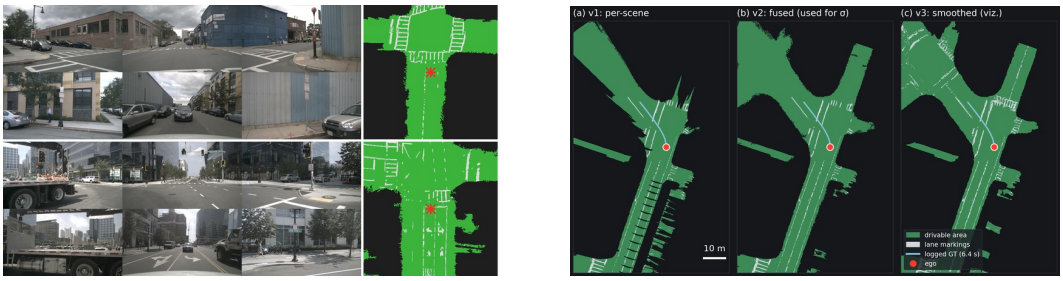


Figure 3: Sensor-derived map. **Left:** six camera views and the fused map (red star = ego, 80 m window) at two intersections. **Right:** one intersection as **(a)** per-scene v1, **(b)** consensus-fused v2 (used for σ) and **(c)** v3 with lane fusion and smoothing (visualization and Appendix A.3.6); green = drivable, white = lanes, blue = logged GT.

3.3 Witness construction and evaluation-set selection

Following the anchor construction in MultiPath (Chai et al., 2019), we cluster training-split trajectories into $K = 256$ representative prototypes. We adapt each prototype to the current ego speed through arc-length re-parameterization under comfort constraints, then verify the resulting trajectory using σ. Trajectories that pass serve as *witnesses*, feasible alternatives beyond the single recorded trajectory. The comfort limits, ego footprint and collision test are specified in Appendix A.1.1.

We identify candidate holes by comparing these witnesses with $N_{\text{det}} = 32$ free-generation samples from the reference planner at each state. A witnessed direction is missing if no reference sample realizes it while satisfying σ, as defined in Eq. 1. Applied to 13,220 training states (15 scenes reserved for held-out evaluation excluded), this yields 6,877 states with at least two witnessed directions, of which 5,689 contain a candidate hole.

Because heading-based labels can mistake a curved road for a branching one, we screen candidate states with a gradient-boosted *validity detector* trained on 1,640 human screening labels. Its 13 features capture witness structure, ego kinematics, missing directions, and road branching without HD-map or maneuver annotations. Branching features count separate drivable exits along circles of radii 14, 18, and 22 m centered on two look-ahead points of the ego log, on the fused map. Clear holes and ambiguous cases are treated as positives, so the detector filters clear non-holes rather than certifying every retained state. At a threshold of 0.6, it retains 3,059 of 5,689 candidate states for evaluation.

3.4 Witness-target augmentation and prefix-scheduled decoding

Training. We select targets from the full candidate pool before detector filtering. For each candidate state–direction pair, we choose up to two witnesses: the medoid and, when a second witness exists, the one farthest from it (98 of 2,933 pairs have a single witness). We discard witnesses whose net heading disagrees with the target direction and cap each scene at 16 targets by round-robin selection over directions. This yields 5,768 targets across 2,933 state–direction pairs at 2,560 states, about 38% of candidate pairs.

Each target is paired with a direction sentence from a fixed bank of 25 phrases per direction, which replaces the record's chain-of-thought text, with text loss restricted to it. We mix these records 1:1 with ground-truth log records (rehearsal). We use Alpamayo-1.5-10B as the vision-language planner and fine-tune it from the off-the-shelf checkpoint, not from the reference planner (training configuration: §4).

Decoding. We propose prefix-scheduled decoding (PSD), which repeats the sample-slot sequence $\{\text{L}, \text{R}, \text{free}, \text{free}\}$. Left and right slots prepend the corresponding direction sentence and free slots use no prefix; the schedule is fixed across scenes, needs no route command, and a budget of N uses its first N slots. We allocate no dedicated straight slot because free generation already produces straight trajectories frequently, as Appendix Table 11 shows. PSD changes only the conditioning of each sample, leaving the trajectory sampler and its flow-step budget unchanged.

4 EXPERIMENTAL SETUP

Evaluation sets. We evaluate recovery on a fixed set of holes identified by the reference planner ($N_{\text{det}} = 32$) and screened by the validity detector. The main evaluation uses a scene-stratified subset of 200 states (254 state–direction pairs); each model generates $N = 128$ samples per state, independent of the detection samples. We additionally evaluate the human-audited gold, validation, held-out, and alternative-witness sets listed in Appendix Table 6; validation trajectories are scored with the official HD map.

Models. We compare off-the-shelf Alpamayo-1.5-10B, its nuScenes (Caesar et al., 2020) GT-finetuned reference, and our witness-augmented model under free generation and PSD. The reference also defines the fixed hole set. Training configurations are detailed in Appendix A.1.3.

Metrics. For a fixed hole pair (x, d), recovery succeeds if at least one of N generated trajectories realizes direction d while satisfying σ:

$$C_N(x, d) = \mathbf{1}\big[\exists i \le N : \sigma(\tau_i) = 1 \wedge D(\tau_i) = d\big]. \tag{2}$$

Closure is the fraction of recovered pairs in the fixed evaluation set $\mathcal{H}$:

$$\text{closure@}N = \frac{1}{|\mathcal{H}|} \sum_{(x,d)\in\mathcal{H}} C_N(x, d). \tag{3}$$

A state missing two directions contributes two pairs. We report closure at $N \in \{8, 16, 32, 64, 128\}$ with scene-bootstrap 95% confidence intervals, the budget for 70% closure (N^*@70%, log-linear interpolation), minADE@8 to the logged trajectory under each row's decoding protocol, and the loop rate.

Data exposure. The training-split sets assess recovery within the training split: 53.1% of main-set pairs and 57.6% of gold-set pairs received a witness target for that state and direction, whereas no validation or held-out pair did. However, the official nuScenes validation split was used to select the reference checkpoint by loss, and 46 validation states were used once to check an earlier configuration; the final model followed the fixed recipe. The 15 held-out scenes were excluded from augmented-planner and detector fitting and from model selection, but their logs contributed to reference-planner training, map fusion, and the prototype bank (Appendix A.2.1).

5 RESULTS

Table 1: Component comparison for Alpamayo-1.5-10B on 200 states (254 pairs). Closure is reported in % under sensor-derived-map σ, with scene-bootstrap 95% CIs. ‡Reference planner whose failures define the evaluation set.

Model	c@8	c@32	c@128 [95% CI]	minADE@8 (m)	N^*@70
Off-the-shelf (Alpamayo-1.5-10B)	13.4	27.2	46.5 [39.0,53.5]	2.39	> 128
+ PSD	19.3	34.3	54.7 [47.3,62.1]	2.24	> 128
GT-finetuned‡	0.0	11.4	23.6 [17.8,29.8]	1.26	> 128
+ PSD	2.4	14.2	28.0 [21.7,34.2]	1.24	> 128
Ours	33.5	64.6	85.0 [79.6,90.1]	1.72	49.4
+ PSD	**44.1**	**71.3**	**90.9 [86.0,95.1]**	1.91	**29.3**

5.1 MAIN RESULTS

As Table 1 shows, witness-target augmentation raises closure@128 from 46.5% to 85.0%, compared with 23.6% for the GT-finetuned reference on the same fixed hole pairs. PSD further raises closure to 90.9% and reduces the estimated budget for 70% closure from 49.4 to 29.3 samples.

Under free generation, the augmented model reaches a minADE@8 of 1.72 m, 0.67 m better than the off-the-shelf model but 0.46 m behind GT-only fine-tuning. PSD adds 0.19 m to this error while improving closure. Figure 4 shows two states where the off-the-shelf planner produces no σ-feasible left turn in 128 samples and Ours + PSD recovers it.

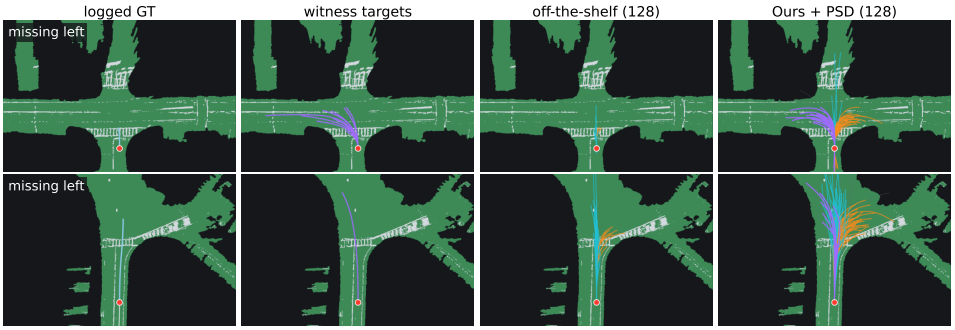


Figure 4: Qualitative recovery on two training-split states with a missing left turn (BEV, ego = red marker, heading up). Columns: logged trajectory (blue) with ego history (white); σ-feasible witness targets for the missing direction (purple); 128 free samples from off-the-shelf Alpamayo-1.5; 128 samples from Ours + PSD. σ-feasible samples are colored by maneuver (L purple, S cyan, R orange). Gray = σ-infeasible.

Table 2: Pair-level closure@128 (%, 95% CIs). Set size = states / pairs. Rows are Ours unless indented. Indented rows are baselines on the preceding set. ‡Reference planner.

Set / model	PSD	Free generation
Train, human-audited gold (502 / 731)	90.0 [86.9,93.1]	85.4 [81.5,88.9]
Train, detector-selected (200 / 254)	90.9 [86.0,95.1]	85.0 [79.6,90.1]
Validation, official-map σ (153 / 176)	96.0 [91.9,99.3]	86.4 [77.0,93.9]
Off-the-shelf	46.6 [35.5,58.4]	38.1 [27.4,49.3]
GT-finetuned‡	21.0 [8.9,34.6]	14.2 [4.2,25.9]
Held-out scenes (66 / 85)	95.3 [88.9,99.2]	96.5 [91.1,100.0]
Off-the-shelf	71.8 [60.2,86.4]	57.6 [39.6,80.5]
GT-finetuned‡	54.1 [40.4,68.2]	56.5 [43.8,70.7]
Map-skeleton witnesses (112 / 138)	90.6 [82.9,97.0]	84.1 [77.0,90.2]

5.2 Recovery across evaluation sets

Recovery remains high across the evaluation sets in Table 2: Ours + PSD closes 96.0% of validation pairs scored with the official HD map, against 46.6% for off-the-shelf Alpamayo and 21.0% for the GT-finetuned reference, and the human-audited gold set closes at 90.0%, while held-out closure meets the pre-registered CI-overlap criterion. Recovery extends to pairs at states without witness-target supervision, and witnessed-direction coverage also improves on validation states selected independently of reference failures, although sample feasibility and logged-direction coverage decline (Appendix A.2).

Table 3: Non-targeted vs. hole-targeted witness training under the same training recipe and target budget.

Set	Decoding	Targets	c@8	c@32	c@128 [95% CI]	N^*@70	minADE@8 (m)
Train 200	PSD	non-targeted	40.9	67.3	85.0 [79.6, 90.1]	37.4	1.81
	PSD	hole-targeted	44.1	71.3	**90.9 [86.0, 95.1]**	29.3	1.91
	free	non-targeted	31.5	57.1	82.7 [76.9, 88.6]	65.1	1.58
	free	hole-targeted	33.5	64.6	85.0 [79.6, 90.1]	49.4	1.72
Validation 153	PSD	non-targeted	60.2	83.5	93.2 [87.5, 98.0]	12.9	2.14
	PSD	hole-targeted	70.5	88.1	**96.0 [91.9, 99.3]**	$\leq$8	2.09
	free	non-targeted	33.0	64.2	82.4 [72.3, 91.6]	44.1	1.38
	free	hole-targeted	44.3	69.9	86.4 [77.0, 93.9]	32.3	1.57
Held-out 66	PSD	non-targeted	42.4	82.4	96.5 [92.4, 100.0]	22.6	1.67
	PSD	hole-targeted	43.5	83.5	95.3 [88.9, 99.2]	15.8	1.57
	free	non-targeted	38.8	81.2	98.8 [95.8, 100.0]	23.4	1.46
	free	hole-targeted	45.9	87.1	96.5 [91.1, 100.0]	19.8	1.40

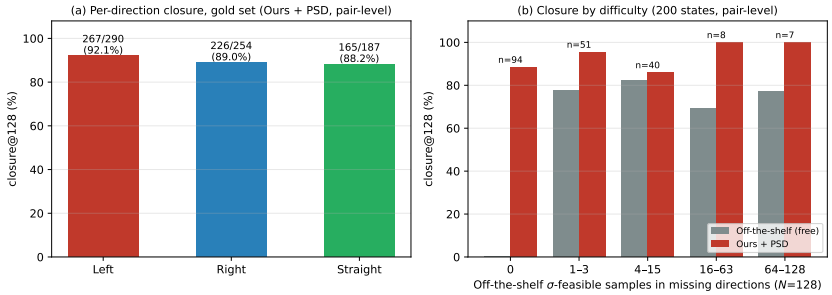


Figure 5: **(a)** Per-direction closure on the gold set (Ours + PSD, N=128). **(b)** Closure by difficulty bin on the 200-state set (state-level bins by the off-the-shelf planner's σ-feasible missing-direction samples in 128 draws).

5.3 Component analysis

Witness augmentation and PSD. PSD increases closure for every model in Table 1. At $N = 8$ its gain is largest for the augmented planner ($+10.6$ pp, against $+5.9$ pp off-the-shelf and $+2.4$ pp for the reference).

Effect of hole targeting. A non-targeted witness control uses the same witness construction and matches initialization, training budget, and target/direction counts. It randomly selects witness targets from multi-option states regardless of whether the reference planner misses the direction (50.5% of its pairs are census holes). Table 3 shows that its closure at $N = 128$ differs from hole-targeted training by less than 6 pp on every set, but validation closure at $N = 8$ is 10–11 pp lower and every set requires more samples to reach 70% closure. Its free-generation minADE@8 is lower on training and validation but not on held-out scenes. Thus, witness-target training accounts for much of the coverage gain, while hole targeting primarily improves sampling efficiency at inference. At the 94 states where the off-the-shelf planner produced no σ-feasible missing-direction sample in 128 draws, Ours + PSD still recovers 88.5% of pairs, as Fig. 5 shows. Prefix compliance and the prior Alpamayo release are reported in Appendix A.4.

Table 4: Lane-direction audit: closure@128 (%) under σ^+ (σ with lane-direction compliance) and wrong-way share among σ-feasible samples (official-map criterion) on the training-split (200 / 254) and validation (153 / 176) sets. Filtered = retrained on the targets kept by the official-map lane filter (an oracle). ‡Reference planner.

		Train 200		Validation 153	
Model (targets)	Decoding	c@128 (σ^+)	wrong-way	c@128 (σ^+)	wrong-way
Off-the-shelf (none)	free	40.6	11.3%	35.2	4.3%
	PSD	48.8	11.9%	43.2	4.9%
GT-finetuned‡ (logs only)	free	19.7	12.9%	12.5	7.3%
	PSD	23.6	12.5%	18.2	7.0%
Ours (all 5,768)	free	77.2	12.6%	84.1	11.1%
	PSD	85.8	14.6%	93.2	14.0%
Filtered (3,573)	free	71.7	12.0%	81.8	10.1%
	PSD	79.9	12.9%	91.5	12.1%

5.4 Lane direction and wrong-way outputs

Adding official-map lane-direction compliance to σ gives σ^+ (Appendix A.3). Coverage gains persist under σ^+, but augmentation and PSD increase the wrong-way share relative to off-the-shelf generation (Table 4). GT-only fine-tuning also raises this share, although less than Ours on validation under both decoding protocols (maneuver breakdown: Appendix A.3.4).

We retrain after removing targets flagged by the official-map criterion, reducing rehearsal to retain the 1:1 ratio. This oracle control alone uses HD-map labels to select training targets. With PSD, σ^+

Table 5: OmniDrive at 6.4 s (200 states / 260 pairs). Free = logged command, PSD = L/R/logged/logged schedule.

Model	c@8	c@32	c@128 [95% CI]	minADE@8 (m)
OmniDrive 6.4 s, GT-only baseline	15.4	20.8	30.0 [23.0,36.7]	1.07
+ PSD (command schedule)	18.8	30.8	41.2 [33.6,49.0]	1.32
OmniDrive 6.4 s + witness targets (Ours)	12.7	25.0	31.9 [24.6,38.6]	1.06
+ PSD (command schedule)	**26.5**	**43.5**	**60.0 [52.5,66.8]**	1.31

closure remains 79.9% on training and 91.5% on validation. Filtering lowers validation wrong-way share from 14.0% to 12.1%, still above both baselines. Thus, filtering does not eliminate wrong-way outputs, and better coverage of lane-consistent alternatives does not imply better sample-level compliance. The smaller training set also prevents isolating the causal contribution of wrong-way targets.

5.5 Experiments on other planner families

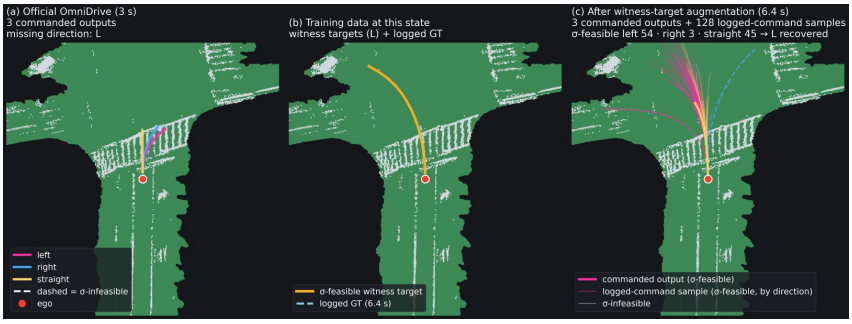


Figure 6: OmniDrive example (missing left turn): **(a)** official 3 s checkpoint under three commands. **(b)** witness target (orange), logged GT (dashed). **(c)** after witness-target augmentation: commanded outputs and 128 logged-command samples.

Diagnosis. OmniDrive (Wang et al., 2025b), a 3D multimodal LLM planner with three-command conditioning, and GenAD (Zheng et al., 2024), a BEV+VAE planner without a language interface, both leave a witnessed direction without a σ-feasible execution on most multi-option states in Appendix Table 13, so holes are not specific to language planners.

Recovery recipe. At 6.4 s with temperature-1.0 sampling, Table 5 shows that witness targets add only 1.9 pp closure@128 under logged-command sampling but 18.8 pp with the command schedule, with nearly unchanged minADE@8 (example in Fig. 6). Unlike Alpamayo PSD, this uses supplied commands and measures command-conditioned coverage. The native 3 s horizon is reported in Appendix A.4.4.

6 Limitations and Conclusion

Evaluation is open-loop and σ omits traffic rules, so wrong-way targets and outputs appear (§5.4). The detector uses one annotator's labels with many ambiguous positives, validation informed reference-checkpoint selection and an earlier configuration check (§4), and the held-out set is one 15-scene draw. PSD is static, and drawing 29–128 samples is offline (Appendix A.1.4).

Cross-scene witnesses identify finite-budget coverage holes and supply recovery targets without HD maps or maneuver annotations. Augmentation raises closure@128 from 46.5% to 85.0% and PSD to 90.9%, at an accuracy cost relative to GT-only fine-tuning; a non-targeted witness control shows that targeting mainly buys sampling efficiency. The gains survive official-map, lane-direction and held-out checks and extend, with command scheduling, to OmniDrive. Wrong-way outputs remain the main open trade-off; rule-aware witness filtering without HD-map annotations and gated prefix slots are the natural next steps.

REPRODUCIBILITY STATEMENT

The appendix specifies the verifier, witness retiming, map fusion, detector and decoding protocols. Evaluation-set specifications, pre-registration records, per-state results and scripts are versioned; maps and witness banks are derived from nuScenes, and detector supervision uses the human screening labels described above.

AI USE STATEMENT

We used generative AI tools to provide feedback on research methodology and experimental design and to assist with interpreting results. We also used these tools to implement parts of the data-processing, training, evaluation, and scoring pipelines, reformat evaluation data, and aggregate results. Additional uses included manuscript translation, organization and editing, literature searches and summarization, reference formatting, and figure preparation. We did not use generative AI tools to generate synthetic datasets or assign the human screening labels used to train the validity detector.

The authors reviewed all AI-assisted work. AI-assisted code was executed, and its outputs were checked against logged results and independent recomputations. Reported numbers were verified against scoring outputs, and citations were checked against the original sources. The authors take responsibility for the final content of this work, including all AI-assisted text, claims, code, and figures.

REFERENCES

Michael Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025.

Sining Ang, Yuguang Yang, Canyu Chen, and Yan Wang. Clover: Closed-loop value estimation and ranking for end-to-end autonomous driving planning. *arXiv preprint arXiv:2605.15120*, 2026.

Mustafa Baniodeh, Kratarth Goel, Scott Ettinger, Carlos Fuertes, Ari Seff, Tim Shen, Cole Gulino, Chenjie Yang, Ghassen Jerfel, Dokook Choe, et al. Scaling laws of motion forecasting and planning–technical report. *arXiv preprint arXiv:2506.08228*, 2025.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11618–11628. IEEE, 2020.

Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. nuPlan: A closed-loop ML-based planning benchmark for autonomous vehicles. *arXiv preprint arXiv:2106.11810*, 2021.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. SAM 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

Yuning Chai, Benjamin Sapp, Mayank Bansal, and Dragomir Anguelov. MultiPath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction. *arXiv preprint arXiv:1910.05449*, 2019.

Changhe Chen, Mozhgan Pourkeshavarz, and Amir Rasouli. CRITERIA: A new benchmarking paradigm for evaluating trajectory prediction models for autonomous driving. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8265–8271. IEEE, 2024.

Keyu Chen, Nanfei Ye, Yida Wang, Wenchao Sun, Danqi Zhao, Hao Cheng, and Sifa Zheng. CRAFT: Counterfactual-to-interactive reinforcement fine-tuning for driving policies. *arXiv preprint arXiv:2605.04470*, 2026.

Alexander Cui, Sergio Casas, Abbas Sadat, Renjie Liao, and Raquel Urtasun. LookOut: Diverse multi-future prediction and planning for self-driving. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16087–16096. IEEE, 2021.

Daniel Dauner, Marcel Hallgarten, Andreas Geiger, and Kashyap Chitta. Parting with misconceptions about learning-based vehicle motion planning. In *Conference on Robot Learning*, pp. 1268–1281. PMLR, 2023.

Fan Ding, Xuewen Luo, Fengze Yang, Bo Yu, HwaHui Tew, Ganesh Krishnasamy, and Junn Yong Loo. KnowDiffuser: A knowledge-guided diffusion planner with LLM reasoning. *arXiv preprint arXiv:2603.10441*, 2026.

Zibin Dong, Yicheng Liu, Yinchuan Li, Hang Zhao, and Jianye Hao. Conditioning matters: Training diffusion policies is faster than you think. *Advances in Neural Information Processing Systems*, 38:119651–119676, 2025.

Catherine Glossop, William Chen, Arjun Bhorkar, Dhruv Shah, and Sergey Levine. CAST: Counterfactual labels improve instruction following in vision-language-action models. *IEEE Robotics and Automation Letters*, 2026.

Maarten Hugenholtz, Anna Meszaros, Jens Kober, and Zlatan Ajanovic. Mode collapse happens: Evaluating critical interactions in joint trajectory prediction models. *arXiv preprint arXiv:2506.23164*, 2025.

Jyh-Jing Hwang, Runsheng Xu, Hubert Lin, Wei-Chih Hung, Jingwei Ji, Kristy Choi, Di Huang, Tong He, Paul Covington, Benjamin Sapp, et al. EMMA: End-to-end multimodal model for autonomous driving. *arXiv preprint arXiv:2410.23262*, 2024.

Bo Jiang, Shaoyu Chen, Hao Gao, Bencheng Liao, Qian Zhang, Wenyu Liu, and Xinggang Wang. VADv2: End-to-end vectorized autonomous driving via probabilistic planning. In *International Conference on Learning Representations*, volume 2026, pp. 68886–68900, 2026.

Bowen Jing, Ruiyang Hao, Weitao Zhou, and Haibao Yu. CounterScene: Counterfactual causal reasoning in generative world models for safety-critical closed-loop evaluation. *arXiv preprint arXiv:2603.21104*, 2026.

Jacky Kwok, Christopher Agia, Rohan Sinha, Matt Foutter, Shulu Li, Ion Stoica, Azalia Mirhoseini, and Marco Pavone. RoboMonkey: Scaling test-time sampling and verification for vision-language-action models. *arXiv preprint arXiv:2506.17811*, 2025.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019.

Zhenxin Li, Kailin Li, Shihao Wang, Shiyi Lan, Zhiding Yu, Yishen Ji, Zhiqi Li, Ziyue Zhu, Jan Kautz, Zuxuan Wu, et al. Hydra-MDP: End-to-end multimodal planning with multi-target hydra-distillation. *arXiv preprint arXiv:2406.06978*, 2024.

Feifei Liu, Haozhe Wang, Zejun Wei, Qirong Lu, Yiyang Wen, Xiaoyu Tang, Jingyan Jiang, and Zhijian He. Beyond ADE and FDE: A comprehensive evaluation framework for safety-critical prediction in multi-agent autonomous driving scenarios. *IFAC-PapersOnLine*, 59(35):442–447, 2025a.

Lin Liu, Guanyi Yu, Ziying Song, Junqiao Li, Caiyan Jia, Feiyang Jia, Peiliang Wu, and Yandan Luo. Beyond imitation: Constraint-aware trajectory generation with flow matching for end-to-end autonomous driving. *arXiv preprint arXiv:2510.26292*, 2025b.

Lin Liu, Caiyan Jia, Guanyi Yu, Ziying Song, JunQiao Li, Feiyang Jia, Peiliang Wu, Xiaoshuai Hao, and Yadan Luo. GuideFlow: Constraint-guided flow matching for planning in end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3719–3728, 2026.

Osama Makansi, Eddy Ilg, Ozgun Cicek, and Thomas Brox. Overcoming limitations of mixture density networks: A sampling and fitting framework for multimodal future prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7144–7153, 2019.

Nicanor Mayumu, Xiaoheng Deng, and Patrick Mukala. Is VLA reasoning faithful? probing safety of chain-of-causation in autonomous driving models. *arXiv preprint arXiv:2605.17268*, 2026.

Rylan Schaeffer, Joshua Kazdan, John Hughes, Jordan Juravsky, Sara Price, Aengus Lynch, Erik Jones, Robert Kirk, Azalia Mirhoseini, and Sanmi Koyejo. How do large language monkeys get their power (laws)? *arXiv preprint arXiv:2502.17578*, 2025.

Ziying Song, Lin Liu, Hongyu Pan, Bencheng Liao, Mingzhe Guo, Lei Yang, Yongchang Zhang, Shaoqing Xu, Caiyan Jia, and Yadan Luo. DIVER: Reinforced diffusion breaks imitation bottlenecks in end-to-end autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Qiao Sun, Weicheng Zheng, Yixin Huang, and Hang Zhao. Is your trajectory displacement safe in long-tail? *arXiv preprint arXiv:2606.16313*, 2026.

Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. DriveVLM: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*, 2024.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *arXiv preprint arXiv:2302.00482*, 2023.

Lingguang Wang, Ömer Şahin Taş, Marlon Steiner, and Christoph Stiller. FlowDrive: Moderated flow matching with data balancing for trajectory planning. *arXiv preprint arXiv:2509.21961*, 2025a.

Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and Jose M. Alvarez. OmniDrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22442–22452. IEEE, 2025b.

Yan Wang, Wenjie Luo, Junjie Bai, Yulong Cao, Tong Che, Ke Chen, Yuxiao Chen, Jenna Diamond, Yifan Ding, Wenhao Ding, et al. Alpamayo-R1: Bridging reasoning and action prediction for generalizable autonomous driving in the long tail. *arXiv preprint arXiv:2511.00088*, 2025c.

Zhongyao Yang, Haoyu Li, Yu Yan, Zhuangxuan Yu, Jiangfeng Nan, and Jinrui Nan. Cognitive dual-process planning for autonomous driving with structured scene knowledge and verifiable reasoning-action consistency. *arXiv preprint arXiv:2607.19194*, 2026.

Michał Zawalski, William Chen, Karl Pertsch, Oier Mees, Chelsea Finn, and Sergey Levine. Robotic control via embodied chain-of-thought reasoning. *arXiv preprint arXiv:2407.08693*, 2024.

Wenzhao Zheng, Ruiqi Song, Xianda Guo, Chenming Zhang, and Long Chen. GenAD: Generative end-to-end autonomous driving. In *European Conference on Computer Vision*, pp. 87–104. Springer, 2024.

A APPENDIX

Section A.1 gives implementation and evaluation details, §A.2 the data-exposure protocol and recovery beyond supervised pairs, §A.3 diagnosis quality and robustness checks, and §A.4 decoding, accuracy and planner comparisons. Pre-registration documents, per-state results and full tables are released with the code.

A.1 IMPLEMENTATION AND EVALUATION DETAILS

Table 6: Evaluation sets. Training-split sets assess recovery within the training split; validation and held-out sets assess recovery on other scenes. Exposure to model selection and offline map/bank construction is described in §4 and Appendix A.2.1.

Set	States	Pairs	Verifier map
Training split, detector-selected	200	254	Sensor-derived v2
Training split, human-audited gold	502	731	Sensor-derived v2
Validation, detector-selected	153	176	Official HD map
Held-out scenes	66	85	Sensor-derived v2
Map-skeleton witnesses	112	138	Sensor-derived v2

A.1.1 VERIFIER AND WITNESS CONSTRUCTION

Feasibility predicate σ. DAC: the ego footprint (4.084×1.945 m, heading-aligned, sampled as a 5×3 point grid per waypoint) must lie entirely inside the drivable mask at every waypoint. NC: at each waypoint, the ego rectangle must not overlap any recorded agent box interpolated at that waypoint's timestamp (a discrete per-waypoint test, not a continuous swept volume; agents are non-reactive; near-stationary ego < 0.5 m/s is excluded to avoid at-fault rear-end artifacts). Direction $D(\tau)$: heading change between the first and last segment, $> +25^\circ$ = L, $< -25^\circ$ = R, else S; trajectories shorter than 3 m are "stop" and never count.

Witness retiming. Each k-means prototype (K=256, 64 waypoints at 0.1 s) is arc-length retimed to the state's ego speed under nuPlan comfort caps (Caesar et al., 2021) ($|a_{\text{lon}}| \leq 2.40$, $|a_{\text{lat}}| \leq 4.89\,\text{m/s}^2$, $|\dot{\psi}| \leq 0.95\,\text{rad/s}$, $|\ddot{\psi}| \leq 1.93\,\text{rad/s}^2$, $|j_{\text{lon}}| \leq 4.13$, $|j| \leq 8.37\,\text{m/s}^3$); a direction is *witnessed* if at least one retimed prototype passes σ.

Direction text. Training direction sentences are drawn from a fixed bank of 25 phrases per direction (5 core sentences $\times$ 5 optional reason clauses, e.g., "Turn left at the intersection ahead", "Keep going straight through the junction as the way ahead is open"); PSD uses the L and R banks as prefixes (the S bank is used only as training text). The phrases are fixed and are not generated from scene observations.

A.1.2 SENSOR-DERIVED MAP CONSTRUCTION

Per-scene drivable masks come from SAM3 road/lane segmentation on the six cameras back-projected to the ground plane (IPM) and accumulated over the scene's timestamps (70 m crop, 0.1 m/px); v2 fuses, at each location, the scene's own observations with a cross-scene vote (kept if vote ratio ≥ 0.4 with ≥ 2 observing scenes, removed if < 0.15 with ≥ 4), followed by component cleaning and hole filling. Fusion uses training-split scenes only (the 15 held-out scenes contribute their own observations, as any scene does, but no augmented-planner or detector fitting touches them); validation scenes have no fused map and are scored with the official HD map. Map quality is evaluated in Appendix A.3.1.

A.1.3 SAMPLING AND EVALUATION CONFIGURATION

Witness-target training starts from the off-the-shelf checkpoint, applies LoRA ($r = 64$, $\alpha = 128$) to the VLM, and fully fine-tunes the expert head with the standard objective for two epochs at a constant learning rate of 10^{-5}. The planner observes the three front camera views plus a narrow front crop and the ego history, and generates a 6.4 s trajectory (64 waypoints at 0.1 s). Except for the R1 comparison in Appendix A.4.2, all Alpamayo results use the public checkpoint

`nvidia/Alpamayo-1.5-10B` (revision `22c33224`) or models fine-tuned from it. Temperature 0.6 and top-p 0.98 (`do_sample`) apply only to chain-of-thought text generation; trajectories are generated separately by the flow-matching expert head using 10 inference steps. Sample k of a state uses seed $1000{+}k$, identical across models. Confidence intervals are percentile intervals from 2,000 scene-level bootstrap resamples with pair-weighted aggregation; paired differences resample the same scenes for both models. $N^*@70\%$ interpolates linearly in $\log_2 N$ between the two budgets that bracket 70% closure, is reported as ≤ 8 when closure@8 already exceeds 70%, and as > 128 when closure@128 does not reach it. The reference planner uses the same trainable modules (LoRA $r{=}64$ on the VLM, full expert head) on logged trajectories only, learning rate $5{\times}10^{-5}$ with cosine decay, best of five epochs by loss on the official nuScenes validation split, which also contains our validation evaluation states (no gradient updates on them).

A.1.4 Compute cost

The recipe changes training data, not the sampler: the fine-tuned planner draws the same N samples with the same number of flow steps as the off-the-shelf model, and PSD only prepends a short direction phrase, so the trajectory sampler and its flow-step budget are identical across rows (end-to-end latency may still differ with prefix and generated-text length). The validation run generated 19,584 samples in 5.0 h on four RTX 3090 GPUs with bf16, about 3.7 GPU-seconds per sample or 7.8 GPU-minutes per 128-sample state; this is an aggregate resource-use estimate, not a measured single-GPU latency comparison. Offline costs, where measured: per-scene BEV accumulation for 626 scenes took about 50 minutes on 28 CPU workers (SAM3 segmentation and cross-scene fusion not timed separately); the reference-planner census (423,040 samples) ran at roughly 5,200 samples per GPU-hour, about 80 GPU-hours; detector feature extraction and fitting take about 1.5 CPU-minutes; the two-epoch fine-tuning run took 29.4 hours on four RTX 3090 GPUs in model-parallel mode (about 118 GPU-hours). Witness verification over 13,220 states is CPU-only and was not timed.

A.2 Data exposure and recovery beyond supervised pairs

A.2.1 Held-out protocol

The 15 reserved scenes were excluded from augmented-planner training (witness targets and rehearsal), detector labels and model selection. Their logs were included in reference-planner training, which was completed on the full nuScenes training split before the scenes were designated, and in map fusion and prototype construction. The held-out evaluation set contains 66 detector-selected states from 11 of the 15 reserved scenes. It was pre-registered and evaluated once under the earlier per-scene map system and model; after the map system was re-locked to the fused v2 map and the planner retrained, the same 66 states were re-registered without re-selection and evaluated once more with the final model. The pre-declared success criterion is that pair-level closure@128 under PSD has a confidence interval overlapping that of the training-split evaluation set in Table 1; both registrations and their results are versioned with the code. As on the main evaluation set, a direction counts as witnessed with at least one σ-feasible witness (183 of the 254 main-set pairs and 76 of the 85 held-out pairs have two or more). Restricting post hoc to those 76 pairs changes little: closure@128 is 94.7% [87.7, 100] for Ours + PSD and 96.1% [90.0, 100] for Ours (free), against 76.3% [63.0, 90.5] and 61.8% [44.6, 83.3] for off-the-shelf Alpamayo (85-pair values: 95.3 / 96.5 vs 71.8 / 57.6).

A.2.2 Closure by training exposure

Splitting the main evaluation set by whether a pair received a witness target for exactly that state and direction (135 pairs), received none (95 pairs), or lies at a state supervised only for another direction (24 pairs): Ours + PSD closes 94.8 / 88.4 / 79.2% of the three groups and Ours (free) 91.9 / 76.8 / 79.2%, against 57.8 / 52.6 / 45.8% for the off-the-shelf planner with PSD, so the gains are similar for pairs at states without witness targets and for directly supervised pairs (35.8 and 37.0 percentage points). On the gold set (421 / 217 / 93 pairs) the split is 91.9 / 87.1 / 88.2%. The lower closure on the 24 pairs supervised only through other directions warrants further analysis.

A.2.3 REFERENCE-INDEPENDENT VALIDATION STATES

To check that coverage gains are not confined to states selected through the reference planner's failures, we fixed 200 validation states before inspecting any model output (seed-fixed, stratified over the 150 validation scenes; 6 overlap the validation hole evaluation set), of which 164 pass the generator's stationary-ego filter. At each state we retime the 256 prototypes and verify them under σ (official HD map DAC $\wedge$ NC), giving witnessed directions independent of any planner. States with at least two witnesses in at least two directions form the evaluation set: 60 states / 137 witnessed pairs (L 40, R 38, S 59); this is stricter than the evaluation-set definition, which requires one witness per direction (183 of the 254 main-set pairs have two or more). *Witnessed-direction coverage@N* is the fraction of pairs with a σ-feasible sample of that direction among the first N samples. Table 7 reports the results; with the looser one-witness criterion (86 states / 197 pairs) the ordering is unchanged (48.2 / 46.2 / 58.9 / 64.0% for off-the-shelf, GT-finetuned, Ours, Ours + PSD). Paired scene-bootstrap differences in coverage@8: Ours $-$ off-the-shelf $+16.1$ [8.6, 23.0], Ours + PSD $-$ off-the-shelf $+20.4$ [11.9, 28.4], Ours + PSD $-$ Ours $+4.4$ [-2.2, 11.2], GT-finetuned $-$ off-the-shelf -0.7 [-7.0, 6.1] pp. The coverage gains persist in a separate run with 32 samples per state on the same states: coverage@32 is 59.9 / 53.3 / 81.0 / 86.9% for off-the-shelf / GT-finetuned / Ours / Ours + PSD, with paired differences Ours $-$ off-the-shelf $+21.2$ [12.7, 29.4], Ours + PSD $-$ off-the-shelf $+27.0$ [17.8, 35.6] and Ours + PSD $-$ Ours $+5.8$ [0.0, 12.3] pp (at N=16: $+8.8$ [2.3, 15.3]). These states were selected independently of reference failures, but their scenes belong to the split used to select the reference epoch (§4).

Table 7: Witnessed-direction coverage on 60 reference-independent validation states (137 pairs; scene-bootstrap CIs). cov@8, per-direction coverage at N=8, feasible (σ-feasible share of samples) and logged (coverage of the logged direction) are from the 8-sample run; cov@16 and cov@32 are from a separate 32-sample run on the same states.

Model	cov@8 [95% CI]	cov@16	cov@32 [95% CI]	L	R	S	feasible	logged
Off-the-shelf (Alpamayo-1.5)	46.0 [40.3, 51.9]	51.1	59.9 [53.8, 66.4]	10.0	36.8	76.3	61.9	85.0
GT-finetuned	45.3 [39.6, 51.5]	51.1	53.3 [46.4, 60.9]	12.5	23.7	81.4	73.3	85.0
Ours (free)	62.0 [53.7, 69.9]	72.3	81.0 [74.3, 87.1]	37.5	63.2	78.0	57.1	80.0
Ours + PSD	**66.4** [56.3, 76.4]	**81.0**	**86.9** [79.4, 93.1]	47.5	71.1	76.3	53.1	75.0

A.3 DIAGNOSIS QUALITY AND ROBUSTNESS

A.3.1 MAP AND TARGET QUALITY

Cross-scene fusion raises drivable-area IoU against the official HD map from 0.381 (per-scene) to 0.535 (v2; the smoothed v3 in Fig. 3 reaches 0.574 but softens branches, so the detector uses v2). Of the 5,768 training targets, 82.1% keep all waypoints inside the official drivable area (L 79.3 / R 79.0 / S 90.9%).

A.3.2 VALIDITY DETECTOR

Specification. Gradient-boosted trees (300 trees, depth 3) on 13 features computed without HD-map or maneuver annotations: $|\mathcal{W}|$, witness counts (L/R/S), $|\Delta\psi|$ of the ego log, windowed-turn ratio, ego speed, missing-direction indicators (L/R/S), and ring-exit statistics on the v2 map at 14/18/22 m (median, max, votes ≥ 3). Human screening labels (clear hole / ambiguous / not-a-hole) were collected on three of four random partitions of the 2,214-state v1 candidate pool: parts 1–2 (1,086 states) and part 3 (554 states), 1,640 in total; part 4 (574 states) was left unlabeled. The held-out protocol fits on parts 1–2 and evaluates on part 3; the deployed detector is fitted on all 1,640 labeled states. Because partitions are random over states, temporally adjacent states of one scene can fall in different parts, so we additionally report a scene-grouped 5-fold cross-validation over all 1,640 labeled states (350 scenes).

Held-out performance. Table 8 and Fig. 7 give the threshold sweep on part 3 (n=554, positives = hole $\cup$ ambiguous, positive rate 68.4%; PR-AUC 0.934). The deployed threshold 0.6 retains 82% of pooled clear-or-ambiguous positives while removing 74% of clear non-holes. Scene-grouped CV

gives PR-AUC 0.900 (state-random CV: 0.927) and, at threshold 0.6, precision 0.83 / recall 0.80 / removal 0.66, i.e., 4, 2 and 8 percentage points below the state-random partition.

Table 8: Validity-detector threshold sweep on the held-out label part (554 states; positives = clear hole ∪ ambiguous). Removal = fraction of clear non-holes rejected. Deployed threshold in bold.

Threshold	0.3	0.4	0.5	**0.6**	0.7	0.8
Precision	0.80	0.82	0.85	**0.87**	0.90	0.93
Recall	0.94	0.91	0.87	**0.82**	0.78	0.69
Removal	0.49	0.57	0.66	**0.74**	0.82	0.89

Clear-hole analysis. Because positives pool clear holes with ambiguous states (49% of positives), we also report the label composition of what the detector selects on part 3 (191 hole / 188 ambiguous / 175 not-a-hole): at threshold 0.6 it selects 357 states = 160 hole + 152 ambiguous + 45 not-a-hole, i.e., recall 0.84 on clear holes and 0.81 on ambiguous states, and precision 0.78 when ambiguous states are set aside (hole vs. not-a-hole). A detector trained and tested on clear holes vs. not-a-hole only (ambiguous excluded; 737 train / 366 test) reaches PR-AUC 0.943 (0.879 scene-grouped) with precision 0.90 / recall 0.79 at threshold 0.6. The gold set used in Table 2 consists exclusively of clear-hole states (502); its 90.0% closure matches the detector-selected 90.9%, so high recovery also holds on clear holes, although this is not an estimate of detector precision.

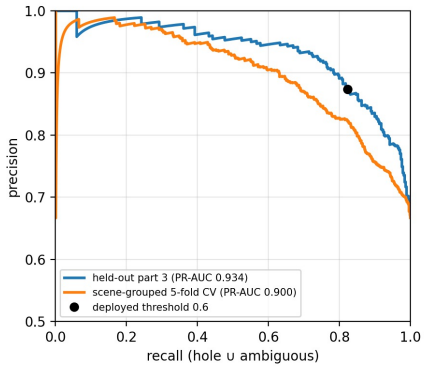


Figure 7: Validity-detector precision–recall: held-out label part (blue) and scene-grouped 5-fold cross-validation (orange); dot = deployed threshold 0.6.

A.3.3 Lane-direction audit

Lane-direction criterion. On evaluation sets where the official HD map provides lane geometry, we add to σ a lane-direction predicate: every fourth waypoint is matched to the nearest lane or lane-connector centerline pose within 3 m (0.5 m discretization), a waypoint is *wrong-way* if its heading deviates from the lane tangent by more than $90°$, and a trajectory passes if fewer than 30% of its matched waypoints are wrong-way (waypoints without a lane within 3 m, e.g., inside intersections, are not counted). $\sigma^+ = \sigma \wedge$ lane-direction. The criterion is calibrated on logged ego trajectories: 0.0% of validation GT left turns (n=72), right turns (n=57) and straight segments (n=678) are flagged, but it is sensitive to intersection geometry: logged GT trajectories on the 200 training-split evaluation states are flagged 5.0% of the time under this default criterion and 1.5% under a more conservative one ($> 120°$ at $\geq 50\%$ of waypoints within 2 m); we report both. The sensor-derived map provides no lane direction, so σ^+ is used only where the official map is available: for the audits below and for the oracle target filter of §5.4. Incorporating lane direction into the witness filter without HD-map annotations remains open.

Witness and target audit. On the 200-state evaluation set, 16.6% of σ-feasible witnesses and 30.7% of the missing-direction witnesses that define holes are wrong-way under the default criterion (L 43.5 / R 40.4 / S 19.6%), and 71 of 254 hole pairs (28.0%) have no lane-direction-consistent witness;

under the conservative criterion the figures are 10.5% and 25/254 (9.8%). On the gold set the default figures are 46.8% and 188/731 (25.7%; GT 9.2%), and the conservative ones 16.6% and 48/731 (6.6%; GT 1.8%). Of the 5,768 witness targets used for training, 38.3% (2,195 of the 5,730 with a determinable lane pose; 38 undetermined targets are kept) are wrong-way under the default criterion (L 47.6 / R 34.9 / S 26.0%) and 15.5% under the conservative one (L 19.1 / R 15.3 / S 9.3%). A lane-direction-aware witness filter would therefore remove roughly 7–28% of the current holes and 16–38% of the training targets (conservative–default criteria).

Recovery under the stricter check. Table 9 gives training-split, validation and held-out closure under σ and σ^+ with the wrong-way share among σ-feasible samples, for the off-the-shelf planner, the GT-finetuned reference, Ours and the model retrained on filtered targets (§5.4); the filtered model's minADE@8 is 1.60 / 1.77 m (free / PSD) on the training split and 1.29 / 1.69 m on validation, against 1.72 / 1.91 and 1.57 / 2.09 m for Ours; on validation, closure@8 under σ^+ is 9.7 / 34.1 / 61.4% for off-the-shelf / Ours / Ours + PSD, and among the samples that first close a missing direction 21.7% (free) and 27.8% (PSD) are wrong-way, concentrated on left turns (31% / 37%). Restricting the training-split evaluation set to the 183 pairs (148 states) with at least one lane-direction-consistent witness gives closure@128 of 50.3 / 88.5 / 92.3% under σ and 47.5 [39.5, 56.2] / 84.2 [78.0, 90.0] / 91.3 [86.2, 95.9]% under σ^+ for off-the-shelf / Ours / Ours + PSD, with N^*@70 (σ) of $>$128 / 31.0 / 23.9, so the ordering and gaps of Table 1 hold on holes that have a lane-direction-consistent alternative from the outset. On the complementary 71 pairs, whose witnesses are all flagged wrong-way, closure@128 under σ is 36.6 / 76.1 / 87.3% for off-the-shelf (free) / Ours (free) / Ours + PSD, lower than on the 183 pairs but of the same order.

Table 9: Closure@128 (%) under σ and σ^+ and wrong-way share among σ-feasible samples on the training-split, validation and held-out sets. The base verifier σ uses the maps listed in Table 6; lane-direction checks use the official HD map. Filtered = retrained without wrong-way targets (§5.4). Brackets: scene-bootstrap 95% CIs where computed.

Set / model	Decoding	c@128 (σ)	c@128 (σ^+)	wrong-way
Train, off-the-shelf	free	46.5	40.6 [33.7, 47.4]	11.3%
	PSD	54.7	48.8	11.9%
Train, GT-finetuned	free	23.6	19.7	12.9%
	PSD	28.0	23.6	12.5%
Train, Ours	free	85.0	77.2 [70.8, 83.2]	12.6%
	PSD	90.9	85.8 [80.2, 91.1]	14.6%
Train, filtered	free	80.3	71.7	12.0%
	PSD	85.8	79.9	12.9%
Validation, off-the-shelf	free	38.1	35.2 [24.2, 47.2]	4.3%
	PSD	46.6	43.2 [30.1, 56.7]	4.9%
Validation, GT-finetuned	free	14.2	12.5	7.3%
	PSD	21.0	18.2	7.0%
Validation, Ours	free	86.4	84.1 [74.1, 92.6]	11.1%
	PSD	96.0	93.2 [86.4, 98.1]	14.0%
Validation, filtered	free	83.5	81.8	10.1%
	PSD	94.3	91.5	12.1%
Held-out, off-the-shelf	free	57.6	52.9	7.4%
	PSD	71.8	65.9	7.7%
Held-out, GT-finetuned	free	56.5	50.6	11.5%
	PSD	54.1	49.4	11.8%
Held-out, Ours	free	96.5	90.6	13.1%
	PSD	95.3	87.1	13.1%
Held-out, filtered	free	95.3	89.4	13.7%
	PSD	96.5	87.1	13.4%

A.3.4 Composition of wrong-way outputs

Table 10 splits the wrong-way share of σ-feasible samples by maneuver class of the sample. On the validation set, turn samples grow from 13% of σ-feasible outputs (off-the-shelf) to 39% (Ours + PSD), and the wrong-way rate within each maneuver class also increases; applying the off-the-

shelf per-class rates to the Ours + PSD maneuver mix would give 8.0% against the observed 14.0%, so the increase is associated with both a larger turn share and higher within-class rates. On the training-split evaluation set the same recomposition gives 17.5% against an observed 14.6%, and the wrong-way rate among left turns falls from 56.0% to 25.7%. This analysis describes how the increase is composed; it does not remove it.

Table 10: Wrong-way share (%) among σ-feasible samples, by maneuver class of the sample (official lane-direction criterion; 128 samples per state). Turn share = fraction of σ-feasible samples that are L or R. Recomposed = composition-only reconstruction: off-the-shelf per-class rates applied to the model's maneuver mix.

Set	Model	turn share	all	L	R	S	recomposed
Val 153	Off-the-shelf (free)	13.4	4.3	13.0	20.3	2.2	4.3
	GT-finetuned (free)	21.1	7.3	22.1	26.6	2.7	5.4
	Ours (free)	25.6	11.1	20.2	26.4	6.6	6.1
	Ours + PSD	38.8	14.0	21.5	24.4	8.1	8.0
Train 200	Off-the-shelf (free)	16.2	11.3	56.0	17.2	6.5	11.3
	GT-finetuned (free)	26.7	12.9	47.6	10.4	6.6	14.8
	Ours (free)	28.8	12.6	26.9	18.3	8.8	14.3
	Ours + PSD	39.4	14.6	25.7	19.8	9.6	17.5
	Filtered targets + PSD	29.0	12.9	27.1	19.4	8.6	15.5

A.3.5 Sensitivity to the direction threshold

We test sensitivity to the default heading-change threshold of $\pm 25^\circ$ by re-labeling witnesses, reference samples, and model samples at $\pm 20^\circ$ and $\pm 35^\circ$ on the same 200 states. For each threshold, we redefine holes as witnessed directions with no σ-feasible reference sample, without reapplying the validity detector. This yields 249, 254, and 202 pairs at $\pm 20^\circ$, $\pm 25^\circ$, and $\pm 35^\circ$, respectively. The default threshold exactly reproduces the original evaluation set.

Closure@128 for off-the-shelf / Ours / Ours + PSD is 49.8 / 88.4 / 91.2% at $\pm 20^\circ$, 46.5 / 85.0 / 90.9% at $\pm 25^\circ$, and 42.1 / 79.7 / 83.7% at $\pm 35^\circ$. The model ordering and substantial coverage gains persist across all three thresholds. Absolute closure is lower at $\pm 35^\circ$, although both direction labels and the evaluated hole pairs change with the threshold.

A.3.6 Alternative witness construction

To check whether recovery extends to holes identified with a different witness construction, we construct witnesses from the map alone on the 1,035 training states that the census had discarded for lacking two witnessed directions. The drivable component connected to the ego on the smoothed v3 map is cleaned with two 3×3 openings and reduced to a one-pixel skeleton (Zhang–Suen thinning); a breadth-first search from the skeleton point nearest the ego yields one path per skeleton endpoint or crop boundary, and dead-end branches shorter than 15 m are dropped. Each path is retimed from the ego speed under a $\pm 2\,\mathrm{m/s}^2$ acceleration limit into 64 waypoints at 0.1 s, labeled L/R/S, and verified with σ on the v2 map. This recovers two or more witnessed directions on 319 states (31%); on 112 of them the reference planner misses a witnessed direction, giving the 138 hole pairs (83 L, 53 R, 2 S) listed in Table 2. The selected model, trained only on prototype witnesses, closes 90.6% [82.9, 97.0] of these pairs with PSD ($N^*@70 = 23.8$). Adding them as further training targets did not yield a statistically significant closure change on any set (McNemar $p \geq 0.29$) but degraded minADE on an 80-state probe set by 16%, so we kept the original model. Because the construction follows centerlines, it does not produce lane changes or alternative lanes on wide roads, and branches broken in the map are not found.

A.4 Decoding, accuracy, and planner comparisons

Schedule choice. Recombining the stored slot samples offline (32 L-prefix and 32 R-prefix samples from the PSD run, the 128 free-generation samples, and the PSD run's 64 free slots) shows how the schedule affects closure. Free generation alone reproduces Tables 1–2 (85.0 / 86.4% at $N{=}128$

Table 11: Prefix→execution confusion for Ours + PSD over the 25,600 evaluation set samples (σ-feasible counts after "/"); the remaining share of each row is stop trajectories (paths shorter than 3 m, excluded from maneuver counting). Off-the-shelf under the same schedule: L→L 24.3%, R→R 16.1%.

Prefix (count)	→ exec L	→ exec R	→ exec S
L (6,400)	**35% / 868**	16% / 400	49% / 1,602
R (6,400)	15% / 357	**47% / 1,058**	38% / 1,193
free (12,800)	16% / 868	20% / 1,156	**62% / 4,442**

on the training split / validation). Interleaving turn prefixes as in $\{\mathrm{L}, \mathrm{R}, \mathrm{free}, \mathrm{free}\}$ raises closure at every evaluated budget in this offline recombination (reconstructed 90.2 / 95.5% at $N{=}128$; the actual PSD run, whose free slots come from the same decoding pass, gives 90.9 / 96.0%), whereas diluting the prefixes to $\{\mathrm{L}, \mathrm{R}, \mathrm{free}{\times}4\}$ lowers it (88.6 / 94.9%). Schedules with a larger turn share gain most at small budgets ($\{\mathrm{L}, \mathrm{R}, \mathrm{free}\}$: 53.1% against 46.9% at $N{=}8$ on the training split, 73.9 against 72.7% on validation) but, with 32 stored samples per turn prefix, support budgets only up to $N{=}96$ ($\{\mathrm{L}, \mathrm{R}, \mathrm{free}\}$) and $N{=}64$ ($\{\mathrm{L}, \mathrm{R}\}$). The same ordering holds for the off-the-shelf and GT-finetuned planners.

A.4.1 Accuracy and path degeneracy

The σ-feasible sample fraction on the training-split evaluation set is 52.3 / 51.4 / 46.7% for off-the-shelf / Ours / Ours + PSD. Loop rate (free generation, path-length/net-displacement > 3, all 25,600 samples per model): Ours 2.39%, GT-finetuned 2.12%, off-the-shelf 0.14%; augmentation adds little over GT fine-tuning, though both fine-tuned models loop more often than the off-the-shelf planner. *Intent-matched minADE@8* (minimum over the first eight samples whose maneuver class equals the logged one; states with no such sample excluded): off-the-shelf 2.59 / 2.47 m (free / PSD, 182 / 181 states), GT-finetuned 1.37 / 1.29 (190 / 189), Ours 2.21 / 2.63 (193 / 195). Restricting to the logged maneuver does not remove the PSD penalty; slot allocation may contribute to it, but the analysis does not isolate its effect, and the included state sets differ across models.

A.4.2 Prior release (Alpamayo-R1-10B)

All fine-tuned Alpamayo models start from the Alpamayo-1.5-10B release, and the off-the-shelf rows use the same checkpoint. Table 12 evaluates the earlier public release, Alpamayo-R1-10B, with the identical protocol. R1 closes more holes than 1.5 but follows L/R prefixes in only 9.2% / 8.0% of slots, so PSD does not help it, whereas the 1.5 release follows prefixes in 24.3% / 16.1% of slots and gains 8 pp from PSD alone (§5.3). Both releases show the same coverage-hole phenomenon; the recipe was evaluated against the release it was trained from.

Table 12: Off-the-shelf releases on the main evaluation set (pair-level closure@128 [95% CI], minADE@8, σ-feasible sample fraction) and on validation (closure@128).

Release	free c@128	PSD c@128	minADE@8 free / PSD	σ-feasible	val free	val PSD
Alpamayo-R1-10B	68.1 [61.0, 74.8]	65.7 [58.4, 72.8]	3.42 / 3.33	44.5%	58.5	56.8
Alpamayo-1.5-10B	46.5 [39.0, 53.5]	54.7 [47.3, 62.1]	2.39 / 2.24	52.3%	38.1	46.6

Table 13: State-level hole rate (a witnessed direction absent) under each planner's native inference protocol and horizon; not an $N{=}128$ sampling comparison.

Planner	Inference protocol	Horizon	Hole rate (states)
GT-finetuned Alpamayo reference	32 free samples	6.4 s	82.7% (5,689 / 6,877)
OmniDrive	native command conditioning	3 s	90.1% (1,855 / 2,059)
GenAD	native command conditioning	3 s	90.3% (1,860 / 2,059)

A.4.3 ADDITIONAL PLANNER DIAGNOSTICS

Using the authors' official checkpoint, GenAD yields a hole rate of 90.3% [88.1, 92.4] (1,860/2,059 states) under the matched 3 s protocol in Appendix Table 13.

A.4.4 WITNESS-TARGET AUGMENTATION ON OMNIDRIVE

Native 3 s horizon. We fine-tune the official checkpoint for two epochs using 2,624 witness targets, 1:1 rehearsal and a frozen vision encoder. OmniDrive uses a numeric command rather than a CoT prefix; under deterministic decoding we evaluate the set of its three commanded outputs. On the paired 3 s evaluation (2,059 states), the hole rate decreases from 89.6% to 73.2%, and commanded-direction failure changes from 78.4% to 45.0% for L, 81.3% to 66.7% for R, and 16.1% to 20.1% for S. The hole-rate reduction is 13.5 percentage points at a $\pm 15^\circ$ direction threshold and 2.2 points at $\pm 10^\circ$, so the estimated benefit depends on how large a heading change qualifies as a turn. The 89.6% baseline re-scores the official-checkpoint outputs from the 90.1% diagnostic run in Table 13, using the v2 map and the paired state set.

Extended 6.4 s horizon. We retrain the model to emit 16 waypoints at 0.4 s intervals: the text trajectory format is extended from 6 to 16 waypoints, the trajectory targets are resampled accordingly, and the model is fine-tuned from the official checkpoint with the tokenizer unchanged. Table 14 separates three training paths: a five-epoch GT-only baseline, a two-epoch witness-augmented model starting from the official 3 s checkpoint, and a two-epoch witness-augmented continuation of the 6.4 s baseline. The second row differs from the baseline in both training data and training duration, so it does not isolate augmentation. The third row shares the trained 6.4 s starting point; relative to the baseline it reduces the hole rate by 3.6 points and left-command failure by 9.0 points, at an ADE increase of 0.12 m. It also receives additional training, so isolating data composition would require an equally long GT-only continuation, which we did not run. Table 5 (main text) compares the first and third rows with temperature-1.0 sampling on 260 pairs, under the logged command and under the L/R/logged/logged schedule; these are command-conditioned results, not command-free generation.

Table 14: OmniDrive at 6.4 s, single-shot ($N=3$ commanded outputs), 6.4 s witnesses, 2,055 multi-option states. Failure means that no σ-feasible trajectory executes the commanded direction. ADE is single-trajectory under the ground-truth command. [§]Differs from the baseline in both data and training budget (2 vs. 5 epochs); the Ours row continues from the 6.4 s baseline checkpoint with two additional epochs.

Model	hole (%)	L fail	R fail	S fail	ADE@6.4 s (m)	ADE@3.2 s (m)
GT-only baseline (5 ep)	96.6	90.3	86.9	48.5	2.15	0.58
Witness targets, official init. (2 ep)[§]	89.9	77.9	87.7	50.0	2.77	0.77
Ours: baseline + targets (2 ep)	93.0	81.3	85.4	50.6	2.27	0.67