TRACE: TASK COMPLETION DETERMINATION FOR AERIAL NAVIGATION WITH SEMANTIC EVIDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Embodied navigation can fail even when an agent reaches the goal region: execution may continue until the agent terminates elsewhere. We study task completion determination in aerial navigation, asking how target-related observations gathered during execution can inform an explicit termination decision. We introduce TRACE, **T**ermination via **R**epeated **A**ssessment of **C**ompletion **E**vidence, a training-free, plug-and-play semantic termination module that attaches to an existing navigation policy without changing its weights. A pretrained vision-language model (VLM) checks the current image against a target description at fixed action-step intervals and reports whether the target is visible. TRACE adds a STOP after consecutive positive reports, while preserving the navigator's own STOP decisions and otherwise leaving movement actions unchanged. The module uses no privileged goal-distance input. Visibility reports guide stopping, while final success is evaluated separately at the terminal position. We evaluate TRACE with frozen OpenFly-Agent and three pretrained VLMs serving as semantic judgers across two simulators. On routes starting outside the goal region, consecutive confirmation yields higher mean success than first-positive stopping across all matched comparisons, while dataset-specific selected configurations also outperform the original navigator with all three VLMs, improving successful termination within the goal region.

1 INTRODUCTION

Vision-language navigation asks an agent to follow a language instruction to its target (Anderson et al., 2018b). Aerial benchmarks extend this task to three-dimensional flight through large scenes (Liu et al., 2023; Gao et al., 2026), including reconstructed cities (Lee et al., 2025) and realistic flight dynamics and guidance (Wang et al., 2025). Reaching the target is only part of the task: success is scored at the terminal position, not the closest point of the flight (Anderson et al., 2018a). Figure 1 shows Baseline passing through the goal region but stopping outside, while TRACE stops inside.

Aerial navigation research has explored representations and action-generation mechanisms for instruction following and goal-region navigation (Liu et al., 2023; Lee et al., 2025; Gao et al., 2026). Approaches include predicting moves from instruction-conditioned video (Zhang et al., 2024), incorporating landmark semantics into top-down maps for action prediction with a large language model (LLM) (Gao et al., 2025), and selecting image-space waypoints with a pretrained vision-language model (VLM) (Hu et al., 2025). Alongside movement, successful navigation requires a decision to end execution within the goal region. We evaluate an added stopping mechanism with OpenFly-Agent's movement policy frozen.

Stopping failures are distinct from failures to reach the target: an agent may stop too early or enter the goal region and then fly through it. On routes that start outside the goal region in our evaluation subset of the OpenFly UE benchmark (Gao et al., 2026), the baseline navigator reaches the goal region in more than half of the episodes but finishes successfully in only 11.52%. More than four out of five episodes that reach the region nevertheless end in failure (Appendix F). Learning when to stop presents a distinct challenge in demonstration-based navigation: a typical trajectory contains many movement actions but only one terminal STOP, creating imbalanced supervision (Xiang et al., 2020). The agent must also assess whether observations support termination, motivating semantic evidence to complement its learned STOP decision.

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.

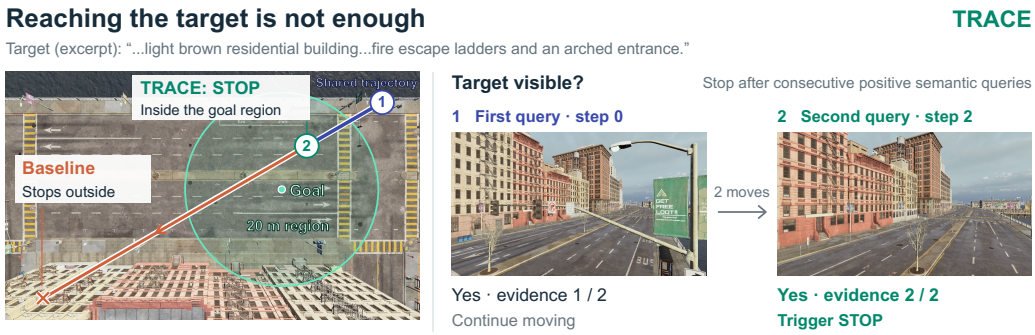

Figure 1: Reaching the target is not enough. On this UE route, Baseline flies through the goal region and stops outside it, while TRACE stops inside after consecutive positive semantic queries.

Existing approaches provide several sources of stopping information. Explicit stopping policies learn termination separately from movement (Xiang et al., 2020; Deng & Xu, 2026); visual success detectors judge task completion from observations and instructions (Du et al., 2023; Duan et al., 2025); and visibility-aware refinement connects target visibility to the terminal position (Shi et al., 2026). Their supervision differs: explicit stopping policies use expert action labels, SuccessVQA fine-tunes a VLM with human reward annotations, and AHA learns failure detection and explanation from procedurally generated manipulation data. We instead study how a pretrained VLM's target-visibility reports from the current RGB image and target description can supply an additional termination signal without task-specific fine-tuning.

We study *task completion determination*: how should target-related observations inform termination? Visibility alone does not establish arrival: a single sighting may be transient or show a target too far away. We introduce TRACE to require persistent support before adding STOP. We use a pretrained vision-language model as a semantic judger, periodically querying the current visual observation for target visibility. These judgments form temporal evidence that supports an additional termination decision when the judger repeatedly reports the target as visible at successive query times (Figure 2). Native STOP retains priority, and movement actions otherwise pass through unchanged. This module requires neither retraining nor privileged goal-distance input.

We evaluate the stopping rule with frozen OpenFly-Agent across UE and AirSim, three pretrained VLMs serving as semantic judgers, and the complete (K, q) hyperparameter grid, where K is the confirmation threshold and q is the query interval. We examine final success, stopping outcomes, movement and query cost, and environment and judger variation. On the evaluated outside-route sets, consecutive confirmation achieves higher mean success rate (SR) than first-positive stopping with the same judger and query interval. Under the selected configurations, the proposed plug-and-play stopping module improves the baseline navigator's success rate by enabling successful termination within the goal region.

Our contributions are:

- We formulate task completion determination and its outcomes, distinguishing target visibility, goal-region reach, issued STOP, and final success with an existing navigation policy whose parameters remain unchanged.
- We introduce a training-free module that converts periodic current-frame visibility judgments from a pretrained VLM into additional STOP decisions through consecutive confirmation, without modifying navigation weights or accessing policy logits or goal distance.
- We systematically evaluate stopping configurations with frozen OpenFly-Agent across two simulators and three VLMs, documenting changes in final success, stopping outcomes, movement, and query cost across the complete (K, q) grid.

2 Related Work

Vision-language navigation. Aerial navigation benchmarks support instruction following through city-scale simulation (Liu et al., 2023), real-city reconstructions with human trajectories and geo-

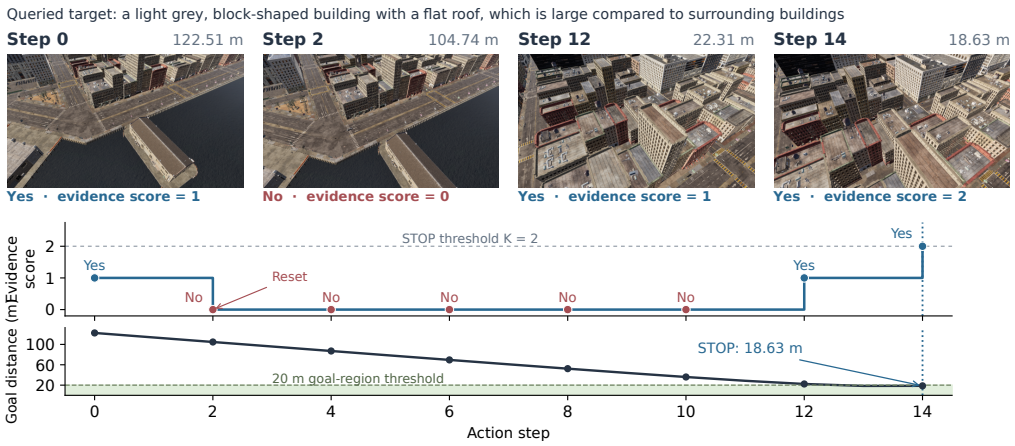

Figure 2: Recorded query sequence and semantic STOP on a UE486 outside route with Qwen at $K = 2, q = 2$, repetition 1. The lower panels show all eight fresh queries. The No at step 2 resets the evidence score, and the Yes at step 14 reaches $K = 2$, triggering STOP at 18.63 meters from the goal. The evidence score is held between queries.

graphic semantic maps (Lee et al., 2025), and flight dynamics with assistant-guided search (Wang et al., 2025). OpenFly provides simulation infrastructure and a keyframe-aware agent (Gao et al., 2026), which we use as the navigator without updating its parameters. In these tasks, visual and linguistic information guides movement. Action selection can use instruction-conditioned video histories (Zhang et al., 2024), top-down landmark maps for LLM prediction (Gao et al., 2025), or pretrained VLM-selected image waypoints converted into 3D movement without navigation training (Hu et al., 2025). In object-goal navigation, vision-language scores and depth-based occupancy maps guide frontier selection (Yokoyama et al., 2024). Beyond immediate action selection, navigation state can be tracked through progress estimates for instruction grounding (Ma et al., 2019) or instruction-progress tracking with landmark memory (Wu et al., 2026). Latent environmental dynamics further support adaptive planning (Yao et al., 2025). RATE-Nav uses exploration estimates and VLM assessments to redirect low-value regional search (Li et al., 2025). These approaches inform movement and search, while success is evaluated at the terminal position (Anderson et al., 2018a). We study how semantic evidence informs termination with the movement policy fixed.

Stopping policies and terminal refinement. Stopping mechanisms differ in how they obtain a termination signal and where they intervene in execution. Learned stopping policies can separate termination from movement: L2Stop uses a Stop Indicator alongside a Direction Decider (Xiang et al., 2020), while DreamFly's LiteStop augments a frozen diffusion policy's STOP with a head trained on policy logits and expert STOP labels (Deng & Xu, 2026). DBFly derives stopping supervision from target proximity and short-horizon motion convergence in expert waypoints for UAV navigation with an initially visible target (Xue et al., 2026). Like LiteStop, TRACE adds a termination signal with the navigation policy fixed; it uses an external pretrained VLM without training a stopping head or accessing policy logits. Intervention can also occur after the original navigator stops: REALM refines position and orientation with a fine-tuned policy and a visibility-aware stopping loss (Shi et al., 2026). TRACE instead intervenes during the original rollout, preserving native STOP and leaving movement actions unchanged. For temporal confirmation, HSGM requires two consecutive high-level planner STOPs to confirm final-subtask completion (Li et al., 2026). TRACE applies consecutive confirmation to target-visibility reports and executes native STOP immediately.

Visual task-completion assessment. REVERIE couples navigation with localization of an object specified by a referring expression (Qi et al., 2020). Learned assessors use task-outcome supervision: SuccessVQA fine-tunes Flamingo on images or short clips with human reward annotations for visual success assessment (Du et al., 2023), while AHA learns manipulation failure explanation, recovery feedback, and subtask verification from procedurally generated failure trajectories (Duan et al., 2025). Our pretrained VLM reports current-frame target visibility without task-specific fine-tuning. We study consecutive reports for termination, evaluating success at the terminal position.

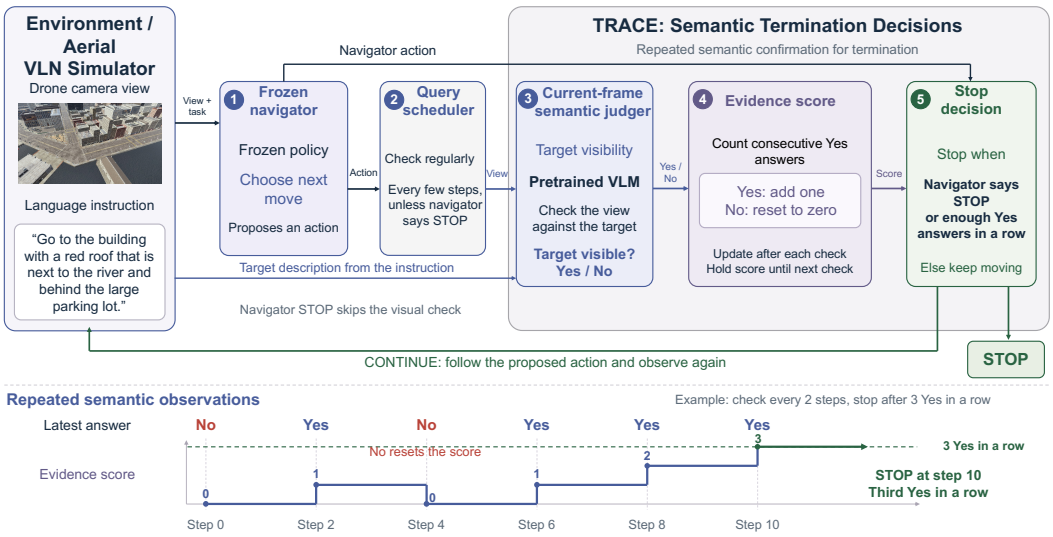

Figure 3: A training-free semantic termination module attached to an existing navigation policy whose parameters remain unchanged. Native STOP takes priority and skips the semantic query. Otherwise, a pretrained VLM serving as the semantic judger checks the current image against the extracted target description every q action steps. Each Yes increments the evidence score, which counts consecutive positive answers; No resets it to zero. The score remains unchanged between queries. Semantic STOP is triggered when a fresh Yes brings the score to K. All other proposed actions execute unchanged. The lower schematic sequence illustrates checks every two steps and termination after three consecutive Yes answers ($q = 2, K = 3$).

3 TRACE: Semantic Evidence for Termination

TRACE augments an existing navigation policy with a training-free termination rule while leaving the policy parameters unchanged. A pretrained VLM serving as the semantic judger reports target visibility from the current image at fixed action-step intervals. An evidence score counts consecutive positive reports and triggers semantic STOP at a confirmation threshold. Figure 3 illustrates the module and query sequence.

3.1 Task completion determination

An agent receives a language instruction x and a visual observation o_t at each action step t. A navigation policy uses the instruction and its available observation history to propose an action a_t from an action space containing movement and STOP. Task completion determination decides whether the available observations support ending execution. We study this decision through an additional termination mechanism that operates on the policy's action proposals.

We distinguish four quantities. *Target visibility* means that the described target appears in the current image. The judger's visibility report is a model prediction of this property and may be incorrect. *Goal-region reach* means that the executed trajectory has entered the goal region, defined by a distance strictly below 20 meters from the goal position in our evaluation. *STOP* is a terminal action issued by the navigator or the added mechanism. *Final success* is the benchmark outcome evaluated at the terminal position. A target can be visible from outside the goal region, a trajectory can enter the region and later leave it, and a STOP can occur at an unsuccessful position. The episode limit can also end execution without a STOP action.

The termination mechanism uses the current image, the instruction, and the proposed action, without access to the simulator's goal position or distance. Goal distance is used only offline to evaluate goal-region reach and final success. Semantic evidence therefore informs the termination decision, while the benchmark independently determines whether the episode succeeds.

3.2 CURRENT-FRAME SEMANTIC JUDGER

We use a pretrained vision-language model as the semantic *judger* without task-specific training or fine-tuning. Let E denote the fixed target-description extraction procedure and J the judger including its response parser. At the jth query, taken at action step t_j, the semantic observation is

$$d = E(x), \qquad y_j = J(o_{t_j}, d) \in \{\text{Yes}, \text{No}\}. \tag{1}$$

Yes indicates that the judger reports the described target as visible in the current image. The prompt uses the target's identifying details to distinguish lookalikes. The judger receives no previous images or answers, leaving cross-query evidence accumulation to the termination mechanism. Its behavior depends on both target-description extraction and visual recognition: an incorrectly extracted destination or omitted identifying modifier can change what the judger is asked to recognize. We instantiate this interface with three pretrained VLMs used as semantic judgers without additional training (Section 4). The prompt, extraction procedure, and parser are described in Appendix A.4.

With steps counted from zero and a fixed episode-wide interval q, the query indicator is

$$m_t = \mathbf{1}[t \bmod q = 0]\, \mathbf{1}[a_t \neq \text{STOP}]. \tag{2}$$

When $m_t = 1$, the current image is queried before executing the proposed movement. The schedule is independent of the answers. Transport retries on the same image count as one logical query.

3.3 EVIDENCE SCORE

A single positive judgment may reflect a transient sighting, a target visible from outside the goal region, or a recognition error. To require persistent semantic support, we maintain an integer evidence score c_j that counts consecutive positive answers. Let $c_0 = 0$ before the first query, with query indices $j = 1, 2, \ldots$. At each fresh valid response, the score is updated as

$$c_j = \begin{cases} c_{j-1} + 1, & y_j = \text{Yes}, \\ 0, & y_j = \text{No}. \end{cases} \tag{3}$$

Between queries, the score remains unchanged; cached responses add no evidence.

The confirmation threshold K specifies the number of consecutive positive answers required to trigger semantic STOP. We call $K > 1$ *temporal-evidence mode* and $K = 1$ *history-free mode*. Here, history refers to semantic evidence retained by the termination module, not the navigator's observation history. With $K = 1$, the first Yes suffices; larger thresholds require positives at successive query times, not at every camera frame. Negative answers reset the evidence score, so an episode may use more than K queries or end without reaching the threshold.

3.4 TERMINATION DECISION

Let $\bar{c}_t$ denote the evidence score after the query at step t when $m_t = 1$. The executed action $\tilde{a}_t$ is

$$\tilde{a}_t = \begin{cases} \text{STOP}, & a_t = \text{STOP}, \\ \text{STOP}, & a_t \neq \text{STOP},\ m_t = 1,\ \bar{c}_t \geq K, \\ a_t, & \text{otherwise}. \end{cases} \tag{4}$$

Native STOP ends execution before any query; semantic STOP ends it immediately upon reaching the threshold. All other movement proposals execute unchanged. The 100-step evaluation limit ends execution with a separate limit reason.

If a run of consecutive positive answers begins at step t^+, the earliest confirmation STOP is

$$t^{\min}_{\text{confirm}} = t^+ + (K - 1)q, \tag{5}$$

provided that subsequent scheduled answers remain positive and neither native STOP nor the step limit intervenes. At a fixed query interval q, increasing K changes both the required response history and the earliest confirmation STOP, while leaving the query schedule and native STOP priority unchanged. Because the navigator continues executing its actions between queries, obtaining further confirmation can change the viewpoints encountered and expose the agent to subsequent stopping

opportunities. For $K > 1$, an isolated positive is insufficient to trigger semantic STOP, while consecutive positives can satisfy the threshold even when the agent remains outside the goal region. Throughout, the mechanism stays fixed and untrained; we only vary the hyperparameters $K, q \in \{1, 2, 3\}$ to probe how the confirmation threshold and query interval affect stopping behavior and cost. Section 4 defines the comparisons and Appendix A gives the full execution protocol.

Figure 2 shows the rule operating on a recorded UE486 route. The first Yes raises the evidence score to one, and the next query returns No, resetting it to zero. Later, two consecutive Yes responses at steps 12 and 14 raise the evidence score to two and trigger semantic STOP. Their offline goal distances are 22.31 and 18.63 meters, respectively: the final confirmation occurs inside the goal region. Appendix B gives the route identity and selection rule. The supplementary material includes an offline HTML visualization with synchronized playback of the Baseline and TRACE executions for this route.

4 EXPERIMENTS

We evaluate the additional stopping rule from three perspectives: final success and paired outcomes relative to Baseline (Section 4.2), mean SR under temporal-evidence mode versus history-free mode with the same judger and query interval (Section 4.3), and variation across environments, judgers, and route initialization groups (Section 4.4).

4.1 EXPERIMENTAL SETUP

Datasets. The OpenFly UE benchmark contains 586 routes across BigCity and SmallCity (Gao et al., 2026). For hyperparameter selection, UE100 contains 31 BigCity and 69 SmallCity routes, approximately preserving the benchmark's city proportions. We evaluate on the remaining UE486 routes, primarily its 382 routes starting outside the goal region. AirSim180 comprises the 180 AirSim routes (Shah et al., 2017) in OpenFly's example evaluation configuration `eval_test.json` (Gao et al., 2026): 30 per scene across six scenes, including 167 starting outside the goal region (Appendix A.2). We call routes starting outside or inside the goal region *outside* or *inside* routes; *all routes* includes both. Outside routes require entering the region for success; inside routes can succeed without moving. Appendix C reports each group and all routes.

Runtime settings. We freeze OpenFly-Agent (Gao et al., 2026); Baseline runs without semantic queries (backbone and scope: Appendix A.1). TRACE uses one of three pretrained VLMs as its semantic judger: LANCE (Fu et al., 2026), Qwen3.8-27B (Qwen; Qwen Team, 2026), and SenseNova-U1.5-8B-MoT (SenseNova; SenseNova, 2026). We sweep all nine combinations of the hyperparameters K (confirmation threshold) and q (query interval), $K, q \in \{1, 2, 3\}$, with three repetitions per judger and combination, plus three Baseline repetitions per environment. Episodes last at most 100 steps, and native STOP takes priority.

Configuration selection. Configurations are selected per dataset and shared across judgers, prioritizing mean SR over $K \in \{2, 3\}$ and $q \in \{1, 2, 3\}$; $K = 1$ is the history-free mode comparison at matched judger and query interval. UE selection uses all inside and outside UE100 development routes and selects $K = 2, q = 2$ for UE486. For AirSim180, we use $K = 3, q = 3$, the most frequently selected pair across leave-one-scene-out folds, for all six scenes and three VLMs. Configuration selection and performance reporting use the same AirSim180 scene collection, so this result is dataset-calibrated rather than an out-of-fold estimate. Appendix A.7 details averaging, judger weights, fold choices, and tie-breaking.

Metrics. We follow OpenFly's evaluation protocol (Gao et al., 2026), including the 20-meter goal-region threshold from AerialVLN (Liu et al., 2023). Success rate (SR) measures the fraction of episodes ending inside the goal region, whereas oracle success rate (OSR) measures the fraction entering it at any point. Success weighted by path length (SPL) additionally accounts for path efficiency (Anderson et al., 2018a); Appendix A.6 specifies our reference-length convention. We also report navigation error (final distance to goal), executed path length, and mean logical queries per episode. Stopping outcomes follow the mutually exclusive categories in Appendix F.

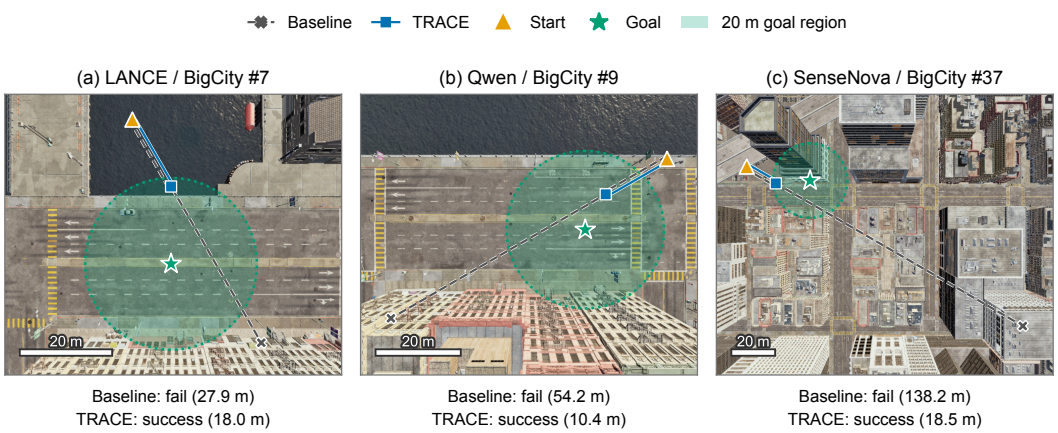

Figure 4: Three UE486 routes where TRACE stops inside the goal region while Baseline does not ($K = q = 2$, repetition 1). Each panel pairs OpenFly-Agent (Baseline, gray dashed) and TRACE (blue) under the indicated judger with the same instruction, start, and target. Crosses and squares mark endpoints; labels give final 3D goal distance. Green circles mark the 20-meter goal region at the shared flight and goal altitude. Top-down backgrounds provide context, not navigation inputs. Appendix B gives the selection rule and failure examples.

4.2 Final success and paired outcomes

The selected configurations increase mean outside-route SR for all three VLMs (Table 1), with paired gains of 5.50–10.65 percentage points on UE486 ($K = 2, q = 2$) and 2.20–5.39 points on dataset-calibrated AirSim180 ($K = 3, q = 3$).

We analyze paired executions where Baseline fails and the method succeeds on the same route and repetition; their movement prefixes may differ. For Qwen on outside UE486, the selected configuration yields 181 failure-to-success transitions and 59 success-to-failure transitions across three repetitions, a net gain of 122 successful executions. Failure-to-success counts exceed success-to-failure counts for all three judgers on both datasets at the selected configurations (Table 2). Among failure-to-success pairs, Baseline reaches the goal region but ends outside in 97.0–97.9% of cases on outside UE486 and 66.7–78.2% on outside AirSim180. Figure 4 shows three UE486 examples.

Table 1: Outside-route means over three repetitions. All rows use frozen OpenFly-Agent; Baseline omits TRACE, and parentheses name TRACE's semantic judger. SR/SPL: %; ΔSR: mean paired gain over Baseline (pp). Each dataset shares one selected (K, q) across judgers. AirSim180 is dataset-calibrated, not out-of-fold.

Method	SR	SPL	ΔSR
UE486 ($K = 2, q = 2$)			
Baseline	11.52	10.82	–
TRACE (LANCE)	17.02	16.98	5.50
TRACE (Qwen)	22.16	21.27	10.65
TRACE (SenseNova)	20.07	19.27	8.55
AirSim180 ($K = 3, q = 3$)			
Baseline	21.56	20.42	–
TRACE (LANCE)	26.95	25.48	5.39
TRACE (Qwen)	23.75	21.85	2.20
TRACE (SenseNova)	23.95	22.01	2.40

Higher SR accompanies lower OSR. For Qwen on outside UE486, SR rises from 11.52% to 22.16%, while OSR falls from 58.81% to 27.75% and the reached-but-failed share from 47.29% to 5.58%. Of 542 Baseline reached-but-failed episodes across three repetitions, 176 succeed under TRACE, 302 end through semantic STOP before ever entering the goal region (premature semantic STOP), 60 remain reached-but-failed, and four end through native STOP or the step limit without entering the region (other non-arrival). Thus, fewer reached-but-failed outcomes reflect both transitions to success and non-arrival in method executions.

4.3 Temporal-evidence mode versus history-free mode

At $K = 1$, a single Yes can trigger semantic STOP before the agent enters the goal region, producing a premature semantic STOP. At the fixed query intervals in Figure 5, raising K to 2 or 3 reduces

Table 2: Baseline-to-TRACE outcome transitions on outside routes: $K = q = 2$ for UE486 and dataset-calibrated $K = q = 3$ for AirSim180. Counts sum three repetitions (3×382 paired executions for UE486; 3×167 for AirSim180). F/S denote failure/success; net gain is F→S minus S→F, and ΔSR is in percentage points. Appendix E gives per-repetition counts.

Outside routes	Judger	F→S	S→F	Net gain	ΔSR
UE486	LANCE	146	83	+63	+5.50
	Qwen	181	59	+122	+10.65
	SenseNova	169	71	+98	+8.55
AirSim180	LANCE	78	51	+27	+5.39
	Qwen	57	46	+11	+2.20
	SenseNova	59	47	+12	+2.40

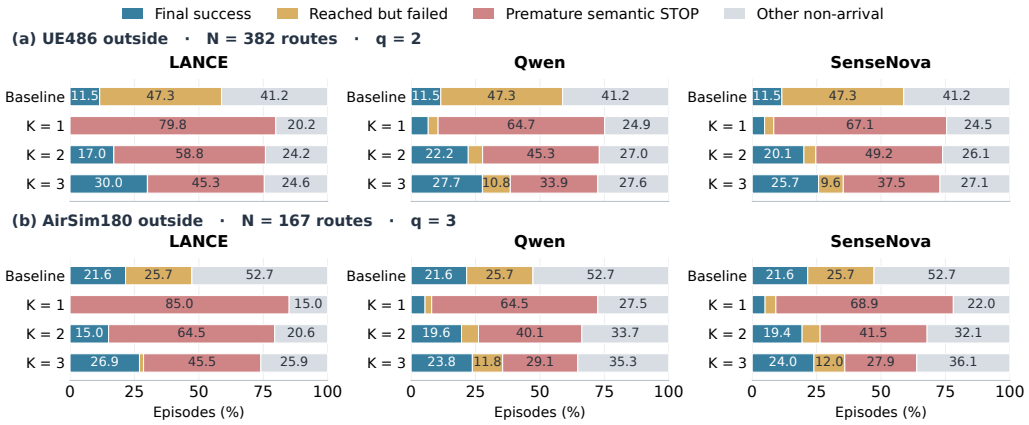

Figure 5: Outside-route outcome composition for UE486 ($N = 382$, $q = 2$) and AirSim180 ($N = 167$, $q = 3$). Panels compare Baseline with $K = 1, 2, 3$ using mean percentages over three repetitions. The four mutually exclusive outcomes are defined in Appendix A.6. Baseline repeats across judgers within each dataset. Segments below 7% retain their widths without labels.

these stops and increases mean SR across all three VLMs. For Qwen on outside UE486 at $q = 2$, the premature semantic STOP rate falls from 64.66% ($K = 1$) to 45.29% ($K = 2$) and 33.94% ($K = 3$), while SR rises from 6.63% to 22.16% and 27.66%, respectively.

Across UE486 and AirSim180 outside routes, all 36 comparisons of $K = 2, 3$ against $K = 1$, matched by dataset, judger, and query interval, yield positive mean SR differences (Appendix E). SR does not increase monotonically with K: for Qwen on outside UE486 at $q = 3$, it rises from 8.90% at $K = 1$ to 24.35% at $K = 2$, then falls to 21.82% at $K = 3$.

Confirmation also increases execution cost in the Qwen outside-UE486 comparison at $q = 2$: raising K from 1 to 2 increases mean path length from 25.25 to 42.37 meters and logical queries from 2.29 to 3.24 per episode (Appendix D). Query counts do not measure wall-clock latency. These comparisons measure the overall effect of temporal-evidence mode relative to history-free mode; they do not isolate subsequent answer content from the additional movement permitted before STOP.

4.4 Variation across environments and judgers

Judger performance varies across environments. Under the dataset-specific selected configurations, Qwen achieves the highest outside-route SR on UE486 (22.16%), whereas LANCE leads on AirSim180 (26.95%). Figure 6 shows how SR varies across the full (K, q) grid; Table 3 summarizes SR, SPL, and Baseline-relative paired transitions.

Configuration sensitivity across datasets. Applying the UE-selected $K = 2, q = 2$ setting within the AirSim180 grid gives mean outside-route SRs of 9.98%, 14.77%, and 16.57% for

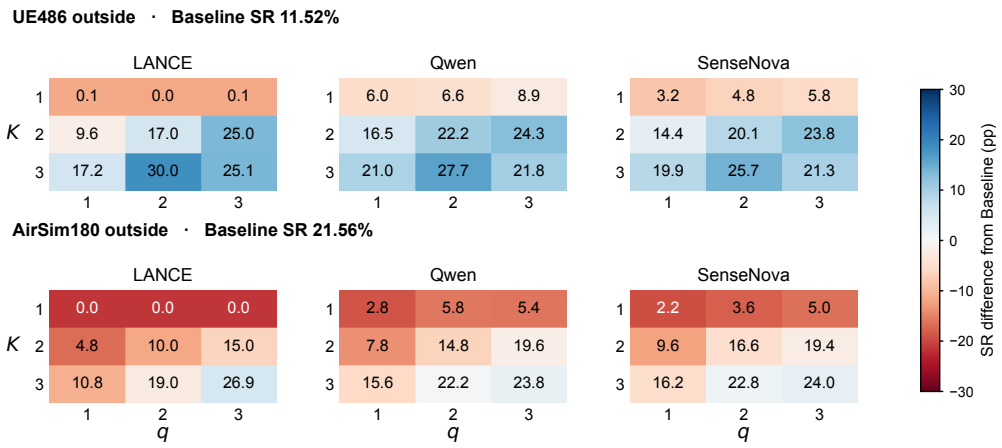

Figure 6: Outside-route SR grid. Each cell prints mean SR (%) over three repetitions; color shows its difference from the same environment's Baseline SR in percentage points. Within each panel, rows are K and columns are q; all three VLMs and all nine settings are shown.

Table 3: Outside-route results across the (K, q) grid. Ranges span configuration means over three repetitions. All columns cover $q \in \{1, 2, 3\}$, with $K = 1$ for the first metric and $K \in \{2, 3\}$ for the rest. SR and SPL are percentages; F→S and S→F count failure-to-success and success-to-failure episodes from Baseline to the method. See Section 4.1 for Baseline results and route counts.

Routes	Judger	$K = 1$ SR	$K = 2, 3$ SR	$K = 2, 3$ SPL	F→S	S→F
UE486 outside	LANCE	0.00–0.09	9.60–30.02	9.60–27.30	27.00–91.33	15.67–34.33
UE486 outside	Qwen	6.02–8.90	16.49–27.66	16.12–24.97	46.67–77.33	12.00–27.67
UE486 outside	SenseNova	3.23–5.85	14.40–25.74	14.25–23.38	39.33–72.33	14.00–28.33
AirSim180 outside	LANCE	0.00–0.00	4.79–26.95	4.79–25.48	5.00–26.00	17.00–33.00
AirSim180 outside	Qwen	2.79–5.79	7.78–23.75	7.61–21.85	8.00–22.33	15.33–31.00
AirSim180 outside	SenseNova	2.20–4.99	9.58–23.95	9.38–22.01	8.67–21.67	15.67–28.67

LANCE, Qwen, and SenseNova, respectively, all below Baseline's 21.56%. Positive mean gains over Baseline on AirSim180 use the separately selected $K = 3, q = 3$ configuration (Appendix C).

5 Limitations

TRACE relies on model-generated visibility judgments, which do not establish geometric arrival; persistent misidentification or distant visibility can still trigger premature semantic STOP. The module preserves native STOP and cannot correct navigation errors through movement. Increasing the confirmation threshold changes both the required evidence and stopping time, so the current comparisons do not isolate the contribution of subsequent semantic responses. Evaluation is limited to frozen OpenFly-Agent in two simulators, with performance depending on configuration and route initialization. The AirSim180 configuration is calibrated on the evaluated scene collection; transfer to unseen scenes, other navigators, and physical flight remains untested.

6 Conclusion

TRACE provides a training-free termination module that converts current-frame VLM visibility judgments into STOP decisions through consecutive confirmation, preserving native STOP priority and leaving movement actions otherwise unchanged. With frozen OpenFly-Agent across two simulators and three VLMs, consecutive confirmation yields higher mean success than first-positive stopping on the evaluated outside-route sets at matched judger and query interval. Dataset-specific selected configurations shared across judgers also improve mean success over Baseline, with the AirSim180 configuration calibrated on the evaluated scene collection.

AI USE STATEMENT

Generative AI tools were utilized to assist with manuscript polishing and boilerplate code implementation for experiments. We also use pretrained vision-language models as experimental components for target-visibility judgments, as described in the method and evaluation protocol. All AI-assisted contributions were rigorously audited: code implementations were manually inspected, unit-tested, and verified for algorithmic correctness by the authors; text revisions were reviewed to ensure precise representation of our methodology. The authors take full responsibility for all content, experimental results, and claims in this paper.

ETHICS STATEMENT

We evaluate the proposed method in simulated aerial navigation environments. Potential real-world applications should account for flight safety and the privacy of people captured in aerial imagery.

REPRODUCIBILITY STATEMENT

Section 3 describes the semantic judger, evidence update, and termination rule. Appendix A details the data splits, model configurations, image preprocessing, exact prompts, response parsing, action timing, evaluation metrics, and hyperparameter selection procedures. Appendix C reports the complete (K, q) hyperparameter grid results, and Appendix G provides results for individual repetitions. We will release the code after the review process.

REFERENCES

Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, and Amir R. Zamir. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*, 2018a. URL https://arxiv.org/abs/1807.06757.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian D. Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 3674–3683. Computer Vision Foundation / IEEE Computer Society, 2018b. doi: 10.1109/CVPR.2018.00387. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Anderson_Vision-and-Language_Navigation_Interpreting_CVPR_2018_paper.html.

Yan Deng and Fei Xu. DreamFly: Causal memory and receding-horizon diffusion planning for aerial vision-language navigation. *CoRR*, abs/2608.12308, 2026. doi: 10.48550/ARXIV.2608.12308. URL https://doi.org/10.48550/arXiv.2608.12308.

Yuqing Du, Ksenia Konyushkova, Misha Denil, Akhil Raju, Jessica Landon, Felix Hill, Nando de Freitas, and Serkan Cabi. Vision-language models as success detectors. In Sarath Chandar, Razvan Pascanu, Hanie Sedghi, and Doina Precup (eds.), *Conference on Lifelong Learning Agents, 22-25 August 2023, McGill University, Montréal, Québec, Canada*, volume 232 of *Proceedings of Machine Learning Research*, pp. 120–136. PMLR, 2023. URL https://proceedings.mlr.press/v232/du23b.html.

Jiafei Duan, Wilbert Pumacay, Nishanth Kumar, Yi Ru Wang, Shulin Tian, Wentao Yuan, Ranjay Krishna, Dieter Fox, Ajay Mandlekar, and Yijie Guo. AHA: A vision-language-model for detecting and reasoning over failures in robotic manipulation. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=JVkdSi7Ekg.

Fengyi Fu, Mengqi Huang, Shaojin Wu, Yunsheng Jiang, Yufei Huo, Hao Li, Yinghang Song, Fei Ding, Jianzhu Guo, Qian He, Zheren Fu, Zhendong Mao, and Yongdong Zhang. Lance: Unified multimodal modeling by multi-task synergy. *CoRR*, abs/2605.18678, 2026. doi: 10.48550/ARXIV.2605.18678. URL https://doi.org/10.48550/arXiv.2605.18678.

Yunpeng Gao, Zhigang Wang, Pengfei Han, Linglin Jing, Dong Wang, and Bin Zhao. Exploring spatial representation to enhance LLM reasoning in aerial vision-language navigation. *arXiv preprint arXiv:2410.08500*, 2025. URL https://arxiv.org/abs/2410.08500v3.

Yunpeng Gao, Chenhui Li, Zhongrui You, Junli Liu, Li Zhen, Pengan CHEN, Qizhi Chen, Zhonghan Tang, Liansheng Wang, Yangpenghui, Yiwen Tang, Yuhang Tang, Shuai Liang, Songyi Zhu, Ziqin Xiong, Yifei Su, Xinyi Ye, Jianan Li, Yan Ding, Dong Wang, Zhigang Wang, Bin Zhao, and Xuelong Li. Openfly: A COMPREHENSIVE PLATFORM FOR AERIAL VISION-LANGUAGE NAVIGATION. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=OKm3w71ymP.

Chih Yao Hu, Yang-Sen Lin, Yuna Lee, Chih-Hai Su, Jie-Ying Lee, Shr-Ruei Tsai, Chin-Yang Lin, Kuan-Wen Chen, Tsung-Wei Ke, and Yu-Lun Liu. See, point, fly: A learning-free vlm framework for universal unmanned aerial navigation. In Joseph Lim, Shuran Song, and Hae-Won Park (eds.), *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 4697–4708. PMLR, 27–30 Sep 2025. URL https://proceedings.mlr.press/v305/hu25e.html.

Jungdae Lee, Taiki Miyanishi, Shuhei Kurita, Koya Sakamoto, Daichi Azuma, Yutaka Matsuo, and Nakamasa Inoue. CityNav: A large-scale dataset for real-world aerial navigation. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pp. 5912–5922. IEEE, 2025. doi: 10.1109/ICCV51701.2025.00559. URL https://doi.org/10.1109/ICCV51701.2025.00559.

Junjie Li, Nan Zhang, Xiaoyang Qu, Kai Lu, Guokuan Li, Jiguang Wan, and Jianzong Wang. Ratenav: Region-aware termination enhancement for zero-shot object navigation with vision-language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, volume ACL 2025 of *Findings of ACL*, pp. 6564–6574. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.FINDINGS-ACL.341. URL https://doi.org/10.18653/v1/2025.findings-acl.341.

Kailing Li, Tianwen Qian, Lijin Yang, Yuqian Fu, Jingyu Gong, Xiaoling Wang, and Liang He. Bridging the 2d-3d gap: A hierarchical semantic-geometric map for vision language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15243–15252, June 2026.

Shubo Liu, Hongsheng Zhang, Yuankai Qi, Peng Wang, Yanning Zhang, and Qi Wu. AerialVLN: Vision-and-language navigation for UAVs. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 15338–15348. IEEE, 2023. doi: 10.1109/ICCV51070.2023.01411. URL https://doi.org/10.1109/ICCV51070.2023.01411.

Chih-Yao Ma, Jiasen Lu, Zuxuan Wu, Ghassan AlRegib, Zsolt Kira, Richard Socher, and Caiming Xiong. Self-monitoring navigation agent via auxiliary progress estimation. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=r1GAsjC5Fm.

Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. REVERIE: remote embodied visual referring expression in real indoor environments. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 9979–9988. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.01000. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Qi_REVERIE_Remote_Embodied_Visual_Referring_Expression_in_Real_Indoor_Environments_CVPR_2020_paper.html.

Qwen Team. Qwen3.8-27B: Official model card. Hugging Face model repository, 2026. URL https://huggingface.co/Qwen/Qwen3.8-27B. Accessed September 15, 2026.

SenseNova. SenseNova-U1.5-8B-MoT: Official model card. Hugging Face model repository, 2026. URL https://huggingface.co/sensenova/SenseNova-U1.5-8B-MoT. Accessed September 15, 2026.

Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. AirSim: High-fidelity visual and physical simulation for autonomous vehicles. In Marco Hutter and Roland Siegwart (eds.), *Field and Service Robotics, Results of the 11th International Conference, FSR 2017, Zurich, Switzerland, 12-15 September 2017*, volume 5 of *Springer Proceedings in Advanced Robotics*, pp. 621–635. Springer, 2017. doi: 10.1007/978-3-319-67361-5_40. URL https://doi.org/10.1007/978-3-319-67361-5_40.

Xiangyu Shi, Ruoxi Yang, Wei Tao, Jiwen Zhang, Yanyuan Qiao, and Qi Wu. From region arrival to instance-level grounding in vision-and-language navigation. *CoRR*, abs/2607.03792, 2026. doi: 10.48550/ARXIV.2607.03792. URL https://doi.org/10.48550/arXiv.2607.03792.

Xiangyu Wang, Donglin Yang, Ziqin Wang, Hohin Kwan, Jinyu Chen, Wenjun Wu, Hongsheng Li, Yue Liao, and Si Liu. Towards realistic UAV vision-language navigation: Platform, benchmark, and methodology. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=rUvCIvI4eB.

Kangyi Wu, Pengna Li, Kailin Lyu, Xi Lin, Lin Zhao, Qingrong He, Jinjun Wang, and Jianyi Liu. Dual-Anchoring: Addressing state drift in vision-language navigation. In *European Conference on Computer Vision*, 2026.

Jiannan Xiang, Xin Wang, and William Yang Wang. Learning to stop: A simple yet effective approach to urban vision-language navigation. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pp. 699–707. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.FINDINGS-EMNLP.62. URL https://doi.org/10.18653/v1/2020.findings-emnlp.62.

Fanfu Xue, En Yu, Bohang Liu, Hongjun Wang, Yang Yang, Xindi Wang, and Jiande Sun. Deliberate before you fly: Vision-guided spatial deliberation for UAV see-and-reach navigation. *CoRR*, abs/2608.04825, 2026. doi: 10.48550/ARXIV.2608.04825. URL https://doi.org/10.48550/arXiv.2608.04825.

Xuan Yao, Junyu Gao, and Changsheng Xu. NavMorph: A self-evolving world model for vision-and-language navigation in continuous environments. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pp. 5536–5546. IEEE, 2025. doi: 10.1109/ICCV51701.2025.00525. URL https://doi.org/10.1109/ICCV51701.2025.00525.

Naoki Yokoyama, Sehoon Ha, Dhruv Batra, Jiuguang Wang, and Bernadette Bucher. VLFM: vision-language frontier maps for zero-shot semantic navigation. In *IEEE International Conference on Robotics and Automation, ICRA 2024, Yokohama, Japan, May 13-17, 2024*, pp. 42–48. IEEE, 2024. doi: 10.1109/ICRA57147.2024.10610712. URL https://doi.org/10.1109/ICRA57147.2024.10610712.

Jiazhao Zhang, Kunyu Wang, Rongtao Xu, Gengze Zhou, Yicong Hong, Xiaomeng Fang, Qi Wu, Zhizheng Zhang, and He Wang. Navid: Video-based VLM plans the next step for vision-and-language navigation. In Dana Kulic, Gentiane Venture, Kostas E. Bekris, and Enrique Coronado (eds.), *Robotics: Science and Systems XX, Delft, The Netherlands, July 15-19, 2024*, 2024. doi: 10.15607/RSS.2024.XX.079. URL https://doi.org/10.15607/RSS.2024.XX.079.

A Evaluation protocol

The evaluation protocol fixes the navigation policy, semantic question, state update, and action rule before comparing confirmation settings. The design crosses two simulators, three pretrained VLMs serving as semantic judgers, and nine hyperparameter combinations (K, q). A run evaluates one experimental condition on all routes in one simulator for one repetition. An episode is a single-route execution; the route-run count totals these episodes across conditions and repetitions. The evaluation comprises 168 runs.

A.1 NAVIGATION BACKBONE SELECTION

Our experiments use a frozen OpenFly-Agent navigator to isolate the effect of the stopping rule: navigation weights and action generation remain fixed across stopping conditions. Our attempts to evaluate additional navigators did not yield a completed cross-navigator comparison: the SPF adaptation encountered simulator communication failures and a different task-termination interface, while our AerialVLN-CMA and AeroAct attempts were limited by the availability of compatible trained navigation checkpoints. The evaluation covers OpenFly-Agent across two environments and three VLMs; transfer to other navigation policies remains to be evaluated.

A.2 DATA SPLITS AND SELECTION INPUTS

UE586 contains 182 BigCity and 404 SmallCity routes. We construct UE100 by selecting 31 BigCity and 69 SmallCity routes, leaving 151 and 335 routes, respectively, in UE486. Routes are grouped by the first component of their original `image_path`. Within each city, we compute the SHA-256 digest of the UTF-8 string formed by concatenating `20260915`, a single newline character, and the unmodified `image_path`. Routes are sorted in ascending lexicographic order by the hexadecimal digest, with `image_path` as the secondary key. The first 31 BigCity and 69 SmallCity routes form UE100; the remaining routes form UE486. Both subsets preserve the original configuration order. The split uses route metadata only, without model outcomes.

UE586 contains 452 outside and 134 inside routes. UE100 contains 70 outside and 30 inside routes; UE486 contains 382 outside and 104 inside routes. AirSim180 contains 167 outside and 13 inside routes across six scenes. Membership is fixed across conditions and repetitions. These summaries overlap and are not independent samples.

AirSim180 source and evaluation manifest. AirSim180 is the AirSim subset of the OpenFly repository's example evaluation configuration `configs/eval_test.json` (Gao et al., 2026). The source configuration contains 240 entries, of which 180 have an `image_path` beginning with `env_airsim_`. These entries already contain 30 routes in each of scenes 16, 18, 23, 26, SH, and GZ; no additional per-scene sampling is applied. OpenFly describes this configuration as an evaluation demonstration, so AirSim180 denotes this collection rather than an asserted reproduction of the paper's official test split.

The evaluation uses the source entries in their original order. The preserved fields are `image_path`, `gpt_instruction`, `action`, `index_list`, `pos`, and `yaw`. The evaluator uses `pos[0]` as the initial position, `yaw[0]` as the initial heading, `pos[-1]` as the evaluation goal position, and `gpt_instruction` as the navigation instruction.

Route initialization is classified offline using the three-dimensional Euclidean distance between `pos[0]` and `pos[-1]`: inside routes have distance below 20 meters, and outside routes have distance at least 20 meters. Table 4 reports the resulting 167 outside and 13 inside routes. The same 180 entries are used across conditions and repetitions.

Table 4: AirSim180 route counts by scene and initial goal-region membership.

Scene	All routes	Outside	Inside
16	30	27	3
18	30	26	4
23	30	28	2
26	30	29	1
SH	30	30	0
GZ	30	27	3
Total	180	167	13

A.3 IMAGE ACQUISITION

Before either Baseline navigation or a semantic-assisted action step, the camera array is checked for empty input and spatial uniformity. Uniformly black frames are included in the rejected class.

Invalid observations are reacquired every 0.5 seconds for up to 30 seconds, without advancing the action index or evidence score. If acquisition still fails, the operational attempt fails and the route chunk is retried; failed attempts and rejected frames are retained.

The original PNG observation is saved at every processed step. A scheduled semantic query encodes the observation as JPEG at quality 85 with contrast enhancement disabled. The query image, prompt, response, and termination-module trace are retained for auditing. The validity check does not reliably identify overexposure, nonuniform brightness abnormalities, or other simulator rendering artifacts. Image retention supports subsequent inspection but does not certify rendering quality.

A.4 SEMANTIC JUDGER INTERFACE

The navigator is the frozen `IPEC-COMMUNITY/openfly-agent-7b` checkpoint (Gao et al., 2026).[1] The judger conditions are LANCE (Fu et al., 2026)[2], Qwen3.8-27B (Qwen Team, 2026), and SenseNova-U1.5-8B-MoT (SenseNova, 2026). Changing the judger leaves the navigator fixed. The model adapters provide the same semantic question, with current-frame image preprocessing and decoding settings held fixed within each judger across all hyperparameter settings (K, q). The generic judger adapter uses temperature zero, a 16-token output limit, and disabled thinking. LANCE uses deterministic decoding, text temperature zero, a 16-token output limit, seed 42, and a 16-token thinking limit in its adapter. Its system prompt is passed through the adapter's system-prompt field. These adapter settings are fixed within each judger across all stopping conditions.

Target-description extraction. Target descriptions are produced by the deterministic text-only function `extract_destination` in `repeated_sensor.py`, version `repeated_semantic_visibility_v1`. The procedure first normalizes whitespace, spacing before punctuation, and spaced hyphens between word characters. Empty instructions are rejected.

It then splits the normalized instruction at predefined punctuation, discourse markers, and comma-delimited movement clauses. Within each resulting clause, it scans destination markers from left to right. A candidate must begin with a supported determiner, contain a noun from the fixed target vocabulary, and pass the exclusions for reference relations and directional phrases. The candidate is truncated at the first matching action suffix. The first valid candidate in each clause is retained, and the candidate from the last qualifying clause is returned.

The procedure retains the remaining destination suffix without noun-head truncation or paraphrasing; modifier retention is therefore determined by the clause-boundary and suffix-truncation rules. If no candidate is accepted, it returns the complete normalized instruction. Extraction receives only the instruction string, without images, route identifiers, goal coordinates, or goal distance.

Listing 1 provides the exact regular expressions and execution order from the frozen implementation. Table 5 shows four input–output pairs from the UE query logs.

Listing 1: Frozen target-description extraction code. The request-construction function and system-prompt constant are omitted from this listing.

```
import re

VERSION = 'repeated_semantic_visibility_v1'
OBJECT = re.compile(r'\b(?:buildings?|skyscrapers?|towers?|structures?|complex(?:es)?|
    landmarks?|monuments?|bridges?|piers?|rooftops?|roofs?|domes?|archways?|entrances?|
    streetlamps?|roads?|streets?|squares?|intersections?|fa[cç]ades?)\b', re.I)
BOUNDARY = re.compile(r'[.;!?]+|\b(?:finally|then|next|afterwards?|following that|again)\b|,\s
    *(?=(?:and\s+)?(?:slightly\s+|gently\s+)?(?:turn|move|go|walk|proceed|head|heading|
    continue)\b)', re.I)
DESTINATION = re.compile(r'\b(?:towards?|to|approach(?:ing)?|reach(?:ing)?|encounter|arrive\s+
    at)\s+', re.I)


def extract_destination(instruction):
    text = re.sub(r'\s+', ' ', str(instruction)).strip()
    text = re.sub(r'\s+([,.;!?])', r'\1', text)
    text = re.sub(r'(?<=\w)\s+-\s+(?=\w)', '-', text)
    if not text:
        raise ValueError('Empty navigation instruction: no target can be defined')
```

[1] Official checkpoint: `https://huggingface.co/IPEC-COMMUNITY/openfly-agent-7b`.

[2] Official repository: `https://github.com/bytedance/Lance`.

```
    candidates = []
    for clause in BOUNDARY.split(text):
        clause = clause.strip(' ,')
        for match in DESTINATION.finditer(clause):
            # Do not mistake a comparison/reference object for the destination.
            prefix = clause[:match.start()]
            if re.search(r'\b(?:compared|relative|adjacent|similar|next|close|opposite)\s*$',
     prefix, re.I):
                continue
            suffix = clause[match.end():].strip(' ,')
            if not re.match(r'^(?:a|an|the|another|this|that)\b', suffix, re.I):
                continue
            if re.match(r'^the\s+(?:left|right)\b', suffix, re.I):
                continue
            if not OBJECT.search(suffix):
                continue
            # Keep all modifiers and relative clauses. Remove only a trailing action.
            suffix = re.split(r',?\s+and\s+(?:stop|move|go|proceed|continue|turn|head)\b',
     suffix, maxsplit=1, flags=re.I)[0].strip(' ,')
            suffix = re.sub(r',?\s+and$', '', suffix, flags=re.I)
            candidates.append(suffix)
            break
    if candidates:
        return {'target': candidates[-1], 'mode': 'last_explicit_destination_clause', 'version
     ': VERSION}
    return {'target': text, 'mode': 'full_instruction_description_fallback', 'version':
     VERSION}
```

Table 5: Recorded target-description extraction examples. Whitespace is normalized for presentation.

Instruction	Extracted description
Ordinary extraction	
Proceed straight towards the large gray brutalist building .	the large gray brutalist building
Modifier retention and normalization	
Proceed straight to the medium - sized , white , historical dome .	the medium-sized, white, historical dome
Multiple destination candidates	
Proceed straight towards the grey open central area surrounded by tall buildings , slightly turn left , and advance forward to the gray building with multiple evenly spaced windows along its facade .	the gray building with multiple evenly spaced windows along its facade
Full-instruction fallback	
The structure , a tall building with light beige coloring accented by white , stands out due to its intricate architectural details when compared to the surrounding shorter and simpler buildings . It is significantly taller than the nearby structures .	The structure, a tall building with light beige coloring accented by white, stands out due to its intricate architectural details when compared to the surrounding shorter and simpler buildings. It is significantly taller than the nearby structures.

The judger input consists of the extracted description and current image, without history images or previous answers. The exact prompts are:

System prompt.

> You are a visual observation module for navigation.
> Judge only the current image against the target description.
> Answer exactly Yes or No, without explanation.

User prompt.

> Target description:
> {target_description}
>
> Is the described target visible in the current image?
> Use the target's identifying details to distinguish it from other objects.
> Answer exactly Yes or No.

The parser accepts a case-insensitive leading Yes or No with a word boundary. It does not accept single-letter aliases or infer a label from an arbitrary substring. A malformed required response invalidates the operational attempt rather than becoming positive or negative evidence. Transport failures may be retried at the same observation and are logged. Response content is not a reason to retry for a different answer. Each successful logical query updates the score once.

A.5 Action timing and termination

At each action step, native STOP is checked first. If it is proposed, the episode ends without querying the judger. Otherwise, at $t = 0, q, 2q, \ldots$, the current image is queried and Eq. 3 is applied. If the evidence score reaches K, the proposed movement is replaced with STOP. Otherwise it is executed unchanged. Non-query steps hold the evidence score and execute the proposed action. The evaluator's translation multiplier (`step_scale`) is fixed at one: basic translations are 3 meters, with forward actions also available at 6 and 9 meters. Turns change yaw by 30°. The maximum episode length is 100 action steps. Native STOP, semantic STOP, and the episode limit are recorded as distinct termination reasons.

A.6 Metrics and outcome categories

The metric set and 20-meter goal-region threshold follow OpenFly's evaluation protocol (Gao et al., 2026), which attributes the threshold to AerialVLN (Liu et al., 2023). The definitions below specify our evaluator's distance, boundary, and termination conventions. For episode i, let S_i indicate final-position success and R_i indicate entry into the 20-meter goal region at any point, including the initial position. Then $\mathrm{SR} = N^{-1} \sum_i S_i$ and $\mathrm{OSR} = N^{-1} \sum_i R_i$. Tables express these proportions as percentages. The set of episodes that reach the goal region can change with the stopping rule. We report SR and OSR separately rather than substituting a conversion ratio for success.

Navigation error (NE) is final goal distance and executed path length (PL) is measured in meters. We use the success-weighted path-efficiency form of SPL (Anderson et al., 2018a):

$$\mathrm{SPL} = \frac{1}{N} \sum_{i=1}^{N} S_i \frac{L_i}{\max(P_i, L_i)}, \tag{6}$$

where L_i is the initial goal distance used as the reference length by the evaluator and P_i is executed path length. The original SPL definition uses shortest-path distance as its reference length (Anderson et al., 2018a); our evaluator uses initial goal distance. When $L_i = 0$, the evaluator assigns the efficiency factor one for $P_i < 10^{-9}$ meters and zero otherwise. Success uses goal distance strictly below 20 meters; outside routes have initial goal distances of at least 20 meters. The grouping of routes by initial position, semantic-query accounting, and outcome categories that distinguish premature semantic STOPs are analysis choices for this study. Query cost is the mean number of logical semantic queries per episode. Each fresh valid response counts once; cached responses and transport retries do not add logical queries. Request attempts count individual network requests, including retries, and are a separate operational measure.

The four mutually exclusive outcome categories are final success, reached but failed, premature semantic STOP, and other non-arrival. Final success means that the episode ends inside the goal region. Reached but failed means that the trajectory enters the goal region but ends outside it. Premature semantic STOP means that semantic STOP ends an episode before it has ever entered the goal region. Other non-arrival means that an episode ends through native STOP or the step limit without ever entering the goal region. Each category's percentage uses all episodes in the reported route group as the denominator; premature semantic STOP excludes native STOP. These categories describe realized outcomes; a truncated route's unobserved continuation is not known. Final success is evaluated independently of the termination reason, so success at the episode limit is not counted as an issued STOP.

For a paired route and repetition, a failure-to-success transition ($\mathrm{F} \to \mathrm{S}$) occurs when Baseline fails and the semantic rule succeeds; a success-to-failure transition ($\mathrm{S} \to \mathrm{F}$) occurs when Baseline succeeds and the semantic rule fails. Both counts use the same eligible route set. Their difference divided by the number of routes equals the paired SR difference for that repetition. The same definition applies when comparing $K = 2$ or $K = 3$ with $K = 1$. Routes are grouped by initial

position before either method acts, using offline goal distance. All goal-distance analyses remain outside the termination-module input.

A.7 Factorial coverage and paired analysis

The grid is $K \in \{1, 2, 3\}$ and $q \in \{1, 2, 3\}$. Three repetitions per judger, grid cell, and simulator give 162 semantic-condition runs. Three Baseline repetitions per simulator add six runs, for a total of 168 runs and 64,344 route-runs across UE586 and AirSim180. Baseline is shared across judger comparisons within an environment and is counted once in coverage. Each condition's repetitions rotate across the three hosts to balance hardware.

UE selection. One common pair is selected from $K \in \{2, 3\}, q \in \{1, 2, 3\}$ on UE100. The score averages each judger's three repetitions and then weights LANCE, Qwen, and SenseNova equally. Higher development SR wins, followed by higher SPL, fewer queries, and smaller K then q; numerical ties use 12-decimal rounding. The selected pair is $K = 2, q = 2$, with mean development SR 40.33%. It is fixed for all three VLMs on UE486. The history-free mode baseline uses $K = 1, q = 2$. Neither UE486 scores nor per-judger winners enter this selector.

AirSim180 configuration selection. We use six-fold leave-one-scene-out analysis to select a common AirSim180 configuration. Each fold ranks the six candidates on five scenes (150 routes), equally weighting Qwen and SenseNova after averaging repetitions, with the same SR/SPL/query/K/q tie-breaking order. Four folds favor $K = 3, q = 3$ (scenes 16, 18, 23, and GZ held out), while two favor $K = 3, q = 2$ (scenes 26 and SH held out). We select $K = 3, q = 3$, the most frequently selected pair, and use it across all six scenes and all three VLMs. Table 20 reports the fold-specific choices.

Configuration selection and performance reporting both use AirSim180. The reported common-configuration comparison evaluates $K = 3, q = 3$ on all 167 outside routes, with $K = 1, q = 3$ as the matched history-free mode baseline; results on all 180 routes are also reported. Although leave-one-scene-out folds are used to obtain configuration choices, evaluating their modal choice on the full collection is not an out-of-fold performance estimate. Such an estimate would evaluate each held-out scene using the configuration selected without that scene and aggregate the resulting held-out evaluations.

Reporting. For each configuration, results are averaged over three repeated runs with fixed decoding settings. Appendix D gives selected metrics, paired effects, and fold choices. Appendix C reports the complete (K, q) hyperparameter grid for all route groups, including development routes.

Aggregate paired means. For each eligible route i, we first average the three repetition-wise paired differences in the binary success indicator:

$$d_i = \frac{1}{3} \sum_{r=1}^{3} \left(S_{i,r}^{\text{method}} - S_{i,r}^{\text{Baseline}} \right). \tag{7}$$

The aggregate difference reported in Table 1 is $\Delta\text{SR} = 100N^{-1} \sum_{i=1}^{N} d_i$ percentage points, where N is the number of eligible routes. Each route has equal weight across the collection, so scenes contribute in proportion to their eligible route counts. The three repetitions are averaged within each route rather than counted as three distinct routes. Differences are computed before rounding the displayed SR values. Comparisons with history-free mode use the same calculation, replacing Baseline with the same judger's $K = 1$ condition at the matched query interval. These means describe the evaluated routes and specified configurations.

A.8 Execution and result validation

Experiments use two RTX 3090 hosts and one A5000 host. AirSim runs one simulator per host because its packaged API port is fixed; navigation and rendering use separate GPUs. Startup selects the NVIDIA Vulkan ICD and waits 20 seconds. Results require complete route coverage and valid executions. Failed operational attempts are retained, and retries use the same scientific configuration.

Replay checks route identity, instruction, initial pose, fresh answers, query indices, evidence-score updates, native STOP priority, executed actions, termination reason, and final metrics. Image and request logs distinguish logical observations from retries.

Metrics and paired comparisons are computed from route-level results using the splits and definitions above. Table 6 summarizes 64,344 route-runs, pooling repetitions and reporting logical queries and termination reasons. Its UE586 counts include the overlapping UE subsets only once.

Table 6: Coverage and termination counts. The evaluation comprises 168 runs and 64,344 route-runs. Total queries counts logical queries summed over all route-runs in each row; termination columns partition route-runs.

Set	Judger	Runs	Route-runs	Total queries	Native	Semantic	Limit
UE586	Baseline	3	1,758	0	1620	0	138
UE586	LANCE	27	15,822	24,003	4208	11614	0
UE586	Qwen	27	15,822	53,738	5515	10215	92
UE586	SenseNova	27	15,822	50,471	5358	10364	100
AirSim180	Baseline	3	540	0	510	0	30
AirSim180	LANCE	27	4,860	7,681	1126	3734	0
AirSim180	Qwen	27	4,860	20,579	1850	2964	46
AirSim180	SenseNova	27	4,860	19,766	1828	2988	44

B PAIRED ROUTE ANALYSIS

Figure 7 illustrates both failure-to-success and success-to-failure transitions on outside UE486 routes at the UE100-selected cell $K = 2, q = 2$ and repetition 1. For each judger and outcome, we select the lexicographically first route ID among paired Baseline/TRACE outcomes whose TRACE termination is semantic STOP. This excludes failure-to-success pairs where TRACE terminates through native STOP. Eligibility counts below refer to that presentation filter, rather than all failure-to-success pairs in the quantitative tables. All paired routes share the instruction, target, initial pose, and repetition; separate executions need not have identical movement prefixes.

Top-down successful stopping examples. Figure 4 uses the same route group, configuration, and repetition. For compact complete-route views, we require Baseline OSR $= 1$, Baseline SR $= 0$, successful semantic STOP, and Baseline path length at most 200 meters. In LANCE, Qwen, and SenseNova order, we choose the lexicographically first eligible route not already shown in an earlier panel. The resulting routes are `001_env_ue_bigcity_005_009_sample002`, `001_env_ue_bigcity_005_009_sample004`, and `007_env_ue_bigcity_035_039_sample002`, respectively. The eligible pools contain 20, 22, and 24 routes before removing duplicates. Their Baseline/TRACE final distances are 27.90/18.00, 54.25/10.39, and 138.21/18.46 meters. Trajectories are projected from the logged metric coordinates using the rendering camera pose and its measured 90° horizontal field of view. Scale bars refer to the goal-altitude plane; all three paired trajectories remain at that altitude.

TRACE (LANCE): successful stopping. The selected route is `001_env_ue_bigcity_005_009_sample002`, from 47 eligible cases. Baseline and TRACE terminal distances are 27.90 and 18.00 meters, respectively.

TRACE (LANCE): failed stopping. The selected route is `000_env_ue_bigcity_000_004_sample000`, from 23 eligible cases. Baseline and TRACE terminal distances are 6.01 and 29.78 meters, respectively.

TRACE (Qwen): successful stopping. The selected route is `001_env_ue_bigcity_005_009_sample002`, from 57 eligible cases. Baseline and TRACE terminal distances are 27.90 and 9.32 meters, respectively.

TRACE (Qwen): failed stopping. The selected route is `000_env_ue_bigcity_000_004_sample000`, from 13 eligible cases. Baseline and TRACE terminal distances are 6.01 and 165.76 meters, respectively.

TRACE (SenseNova): successful stopping. The selected route is `001_env_ue_bigcity_005_009_sample004`, from 56 eligible cases. Baseline and TRACE terminal distances are 54.25 and 10.39 meters, respectively.

TRACE (SenseNova): failed stopping. The selected route is `000_env_ue_bigcity_000_004_sample003`, from 16 eligible cases. Baseline and TRACE terminal distances are 6.58 and 38.63 meters, respectively.

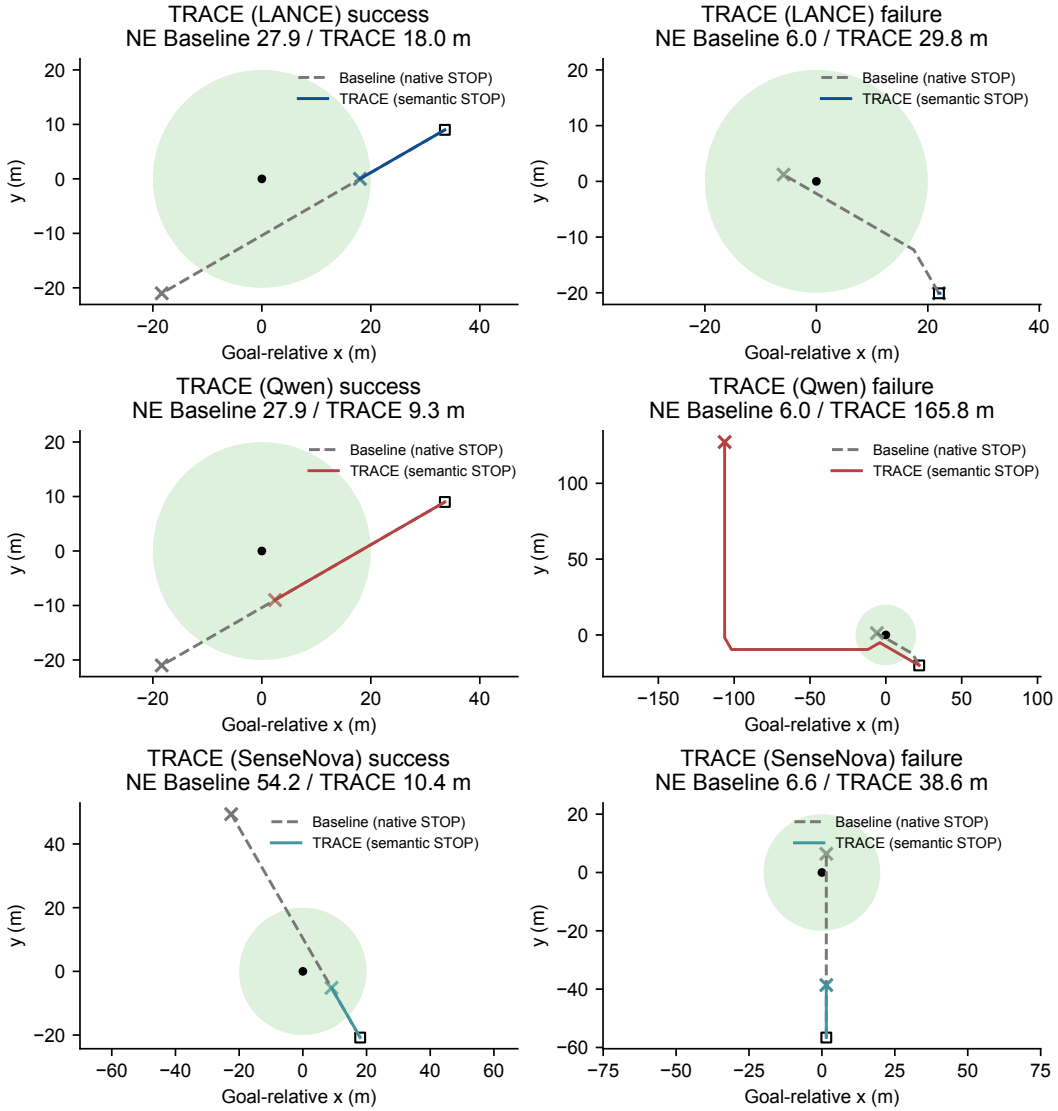

Figure 7: Paired routes at $K = 2, q = 2$, UE486 outside, repetition 1. Each TRACE judger configuration has a failure-to-success pair (left) and a success-to-failure pair (right), selected by the stated rule. Solid and dashed paths show paired executions; endpoints and termination reasons are annotated in the panels. Coordinates are goal-centered horizontal projections. The circle projects the 20-meter goal region; actual success uses three-dimensional distance, including altitude. Distances and outcomes come from the logged trajectories, not from visual target presence.

B.1 A FRESH-QUERY SEQUENCE WITH AN EVIDENCE RESET

For the recorded query sequence in Figure 2 and the timeline in Figure 8, we select the first eligible Qwen route at $K = 2, q = 2$, repetition 1, with a Yes-to-No reversal followed by a successful semantic STOP: `053_env_ue_smallcity_080_084_sample002`. Its eight queries occur at steps 0, 2, 4, 6, 8, 10, 12, and 14, with answers Yes, No, No, No, No, No, Yes, Yes. The evidence score is 1 after the first query, resets to 0 at step 2, and reaches 1 and 2 at steps 12 and 14. The first positive occurs 122.51 meters from the goal; the final two queries occur at 22.31 and 18.63 meters. Thus support begins outside the goal region and the second consecutive positive triggers STOP inside it. This illustrates the mechanism, not visibility accuracy or its frequency.

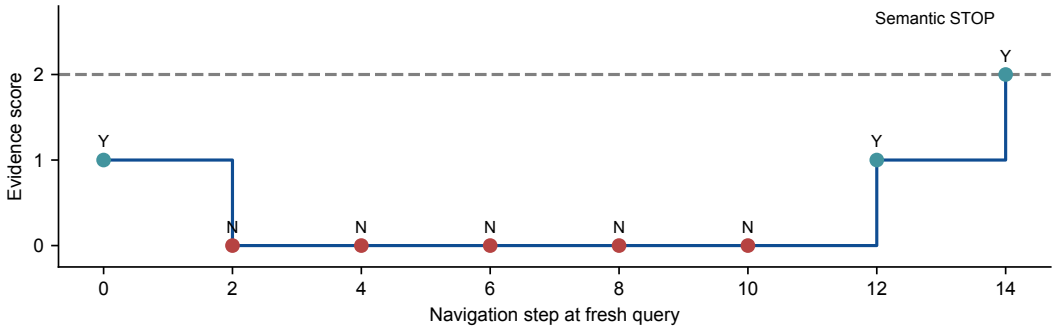

Figure 8: Periodic-query timeline for the specified Qwen route. Markers show fresh answers and the evidence score after each update; non-query steps hold the score. The No at step 2 resets the evidence score, and the Yes at step 14 reaches $K = 2$ and triggers semantic STOP. Offline distance provides the separate final-success evaluation and is never a judger input.

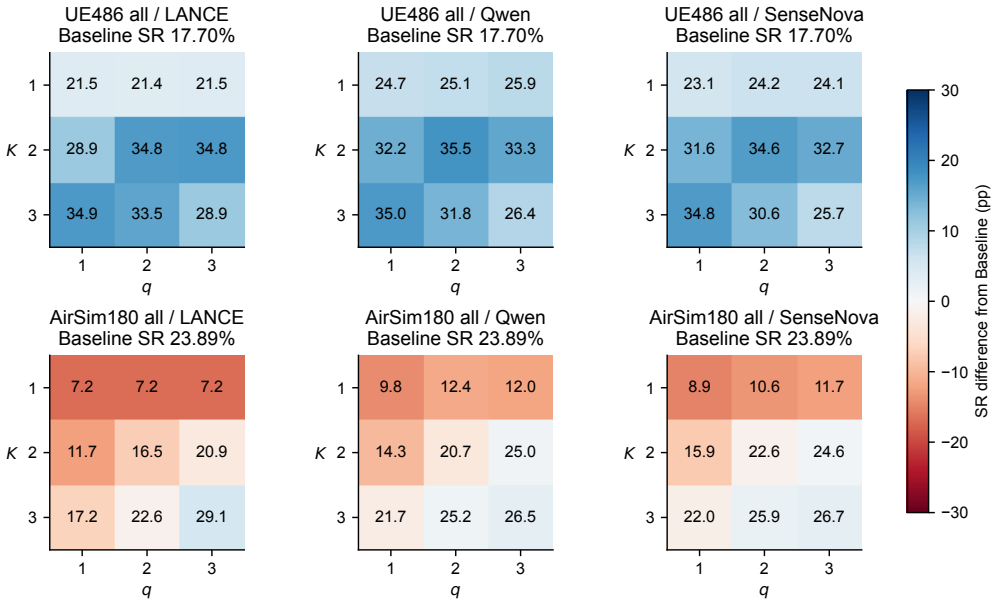

Figure 9: SR over the complete (K, q) hyperparameter grid on all UE486 ($N = 486$) and AirSim180 ($N = 180$) routes, including those that start inside the goal region. Cell values are mean SR in percent; colors are percentage-point differences from the corresponding Baseline mean, not significance. The outside counterpart is Figure 6.

C COMPLETE EXPERIMENTAL RESULTS

All tables and plots use the evaluation protocol and UE100/UE486 split defined in Appendix A. UE586, UE486, UE100, and AirSim180 each have summaries for all routes, outside routes, and inside routes. These 12 summaries overlap and must not be added as independent samples. Each contains Baseline and all 27 combinations of judger and hyperparameter setting (K, q). Values are means of three repetitions, with all grid settings retained alongside the separate selection analysis. Figure 9 shows SR over the complete (K, q) hyperparameter grid on all UE486 and AirSim180 routes, including those that start inside the goal region.

Table 7: UE100, all 100 routes: means of three repetitions. SR, SPL and OSR are percentages. Navigation error (NE) and executed path length (PL) are measured in meters. Queries reports the mean number of logical queries per episode. The hyperparameters K and q do not apply to Baseline. Premature semantic STOP is the percentage of episodes terminated by semantic STOP before ever entering the goal region; native STOP is excluded.

Judger	K	q	SR	SPL	OSR	NE	PL	Queries	Premature semantic STOP
Baseline	–	–	22.00	20.25	72.33	185.53	231.51	0.00	0.00
LANCE	1	1	30.00	30.00	30.00	54.55	0.00	0.81	58.33
LANCE	1	2	30.00	30.00	30.00	54.55	0.00	0.80	58.33
LANCE	1	3	30.00	30.00	30.00	54.55	0.00	0.80	58.00
LANCE	2	1	32.67	26.58	32.67	49.21	6.71	1.60	52.67
LANCE	2	2	41.00	30.43	41.00	45.58	13.36	1.57	43.67
LANCE	2	3	36.67	31.73	47.33	43.14	19.67	1.53	36.33
LANCE	3	1	41.67	31.00	41.67	45.60	13.37	2.34	42.00
LANCE	3	2	33.33	31.21	51.33	41.06	26.02	2.27	33.00
LANCE	3	3	34.00	29.78	58.00	39.98	37.18	2.18	23.33
Qwen	1	1	31.00	29.46	33.33	60.28	22.47	3.33	52.00
Qwen	1	2	31.67	29.03	35.00	60.43	27.29	2.33	50.00
Qwen	1	3	33.33	31.57	37.00	56.86	27.35	1.84	48.00
Qwen	2	1	34.67	26.84	38.33	56.97	32.96	4.57	45.33
Qwen	2	2	41.00	32.40	49.33	54.27	40.53	3.11	33.67
Qwen	2	3	34.67	32.33	55.00	57.94	55.30	2.87	26.67
Qwen	3	1	38.67	29.60	46.33	57.95	42.89	6.01	36.00
Qwen	3	2	36.33	34.15	58.00	57.74	60.72	4.21	23.33
Qwen	3	3	32.00	27.85	62.67	71.96	84.93	3.99	18.67
SenseNova	1	1	31.67	30.70	37.67	61.90	26.41	3.82	49.00
SenseNova	1	2	33.33	31.11	38.33	60.54	28.58	2.48	48.67
SenseNova	1	3	29.67	29.00	39.00	63.99	33.92	2.10	48.33
SenseNova	2	1	35.00	28.34	42.00	65.24	41.52	5.60	43.33
SenseNova	2	2	39.00	29.62	49.00	62.69	52.03	3.78	33.33
SenseNova	2	3	36.33	32.59	55.00	65.68	63.96	3.21	27.33
SenseNova	3	1	37.00	28.16	47.67	67.88	54.67	7.22	34.00
SenseNova	3	2	31.67	29.85	57.00	66.95	68.85	4.68	23.67
SenseNova	3	3	30.67	27.07	63.67	77.64	89.05	4.14	15.33

Table 8: UE100, 30 inside routes: means of three repetitions. SR, SPL and OSR are percentages. Navigation error (NE) and executed path length (PL) are measured in meters. Queries reports the mean number of logical queries per episode. The hyperparameters K and q do not apply to Baseline. Premature semantic STOP is the percentage of episodes terminated by semantic STOP before ever entering the goal region; native STOP is excluded.

Judger	K	q	SR	SPL	OSR	NE	PL	Queries	Premature semantic STOP
Baseline	–	–	41.11	36.50	100.00	138.60	162.27	0.00	0.00
LANCE	1	1	100.00	100.00	100.00	9.59	0.00	0.74	0.00
LANCE	1	2	100.00	100.00	100.00	9.59	0.00	0.73	0.00

Continued

Judger	K	q	SR	SPL	OSR	NE	PL	Queries	Premature semantic STOP
LANCE	1	3	100.00	100.00	100.00	9.59	0.00	0.72	0.00
LANCE	2	1	100.00	79.71	100.00	6.64	7.03	1.57	0.00
LANCE	2	2	100.00	64.78	100.00	9.65	13.33	1.51	0.00
LANCE	2	3	64.44	48.22	100.00	14.96	19.53	1.43	0.00
LANCE	3	1	100.00	64.43	100.00	9.67	13.50	2.27	0.00
LANCE	3	2	40.00	36.29	100.00	20.15	24.97	2.11	0.00
LANCE	3	3	40.00	35.89	100.00	28.56	33.17	1.98	0.00
Qwen	1	1	95.56	90.90	100.00	27.58	22.73	3.30	0.00
Qwen	1	2	93.33	85.71	100.00	29.91	25.60	2.18	0.00
Qwen	1	3	91.11	86.32	100.00	25.16	19.97	1.50	0.00
Qwen	2	1	92.22	67.39	100.00	22.74	24.70	3.52	0.00
Qwen	2	2	82.22	54.12	100.00	35.65	39.70	2.98	0.00
Qwen	2	3	42.22	37.10	100.00	53.07	58.67	2.89	0.00
Qwen	3	1	81.11	51.89	100.00	36.44	40.77	5.27	0.00
Qwen	3	2	42.22	38.94	100.00	48.39	52.53	3.59	0.00
Qwen	3	3	40.00	34.88	100.00	84.45	91.20	4.10	0.00
SenseNova	1	1	93.33	90.76	100.00	29.16	23.27	3.30	0.00
SenseNova	1	2	95.56	88.54	100.00	28.89	23.43	2.06	0.00
SenseNova	1	3	91.11	88.90	100.00	29.54	23.70	1.63	0.00
SenseNova	2	1	96.67	74.91	100.00	26.31	27.17	3.80	0.00
SenseNova	2	2	88.89	58.24	100.00	27.37	31.03	2.48	0.00
SenseNova	2	3	56.67	45.75	100.00	41.18	45.50	2.37	0.00
SenseNova	3	1	85.56	56.66	100.00	32.51	36.33	4.83	0.00
SenseNova	3	2	38.89	36.58	100.00	47.55	52.50	3.63	0.00
SenseNova	3	3	38.89	35.07	100.00	58.93	63.80	3.08	0.00

Table 9: UE100, 70 outside routes: means of three repetitions. SR, SPL and OSR are percentages. Navigation error (NE) and executed path length (PL) are measured in meters. Queries reports the mean number of logical queries per episode. The hyperparameters K and q do not apply to Baseline. Premature semantic STOP is the percentage of episodes terminated by semantic STOP before ever entering the goal region; native STOP is excluded.

Judger	K	q	SR	SPL	OSR	NE	PL	Queries	Premature semantic STOP
Baseline	–	–	13.81	13.28	60.48	205.65	261.19	0.00	0.00
LANCE	1	1	0.00	0.00	0.00	73.82	0.00	0.83	83.33
LANCE	1	2	0.00	0.00	0.00	73.82	0.00	0.83	83.33
LANCE	1	3	0.00	0.00	0.00	73.82	0.00	0.83	82.86
LANCE	2	1	3.81	3.81	3.81	67.46	6.57	1.61	75.24
LANCE	2	2	15.71	15.71	15.71	60.97	13.37	1.59	62.38
LANCE	2	3	24.76	24.66	24.76	55.21	19.73	1.57	51.90
LANCE	3	1	16.67	16.67	16.67	61.00	13.31	2.37	60.00
LANCE	3	2	30.48	29.03	30.48	50.02	26.47	2.33	47.14
LANCE	3	3	31.43	27.17	40.00	44.87	38.90	2.27	33.33
Qwen	1	1	3.33	3.13	4.76	74.30	22.36	3.35	74.29
Qwen	1	2	5.24	4.74	7.14	73.52	28.01	2.40	71.43
Qwen	1	3	8.57	8.11	10.00	70.45	30.51	1.99	68.57
Qwen	2	1	10.00	9.46	11.90	71.64	36.50	5.01	64.76
Qwen	2	2	23.33	23.10	27.62	62.24	40.89	3.17	48.10
Qwen	2	3	31.43	30.28	35.71	60.03	53.86	2.87	38.10
Qwen	3	1	20.48	20.04	23.33	67.17	43.80	6.32	51.43
Qwen	3	2	33.81	32.10	40.00	61.75	64.23	4.48	33.33
Qwen	3	3	28.57	24.83	46.67	66.60	82.24	3.94	26.67
SenseNova	1	1	5.24	4.96	10.95	75.93	27.76	4.05	70.00
SenseNova	1	2	6.67	6.50	11.90	74.11	30.79	2.66	69.52
SenseNova	1	3	3.33	3.33	12.86	78.75	38.30	2.30	69.05
SenseNova	2	1	8.57	8.38	17.14	81.93	47.67	6.38	61.90

Continued

Judger	K	q	SR	SPL	OSR	NE	PL	Queries	Premature semantic STOP
SenseNova	2	2	17.62	17.36	27.14	77.82	61.03	4.34	47.62
SenseNova	2	3	27.62	26.95	35.71	76.17	71.87	3.58	39.05
SenseNova	3	1	16.19	15.95	25.24	83.04	62.53	8.24	48.57
SenseNova	3	2	28.57	26.97	38.57	75.27	75.86	5.12	33.81
SenseNova	3	3	27.14	23.64	48.10	85.65	99.87	4.59	21.90

Table 10: UE486, all 486 routes: means of three repetitions. SR, SPL and OSR are percentages. Navigation error (NE) and executed path length (PL) are measured in meters. Queries reports the mean number of logical queries per episode. The hyperparameters K and q do not apply to Baseline. Premature semantic STOP is the percentage of episodes terminated by semantic STOP before ever entering the goal region; native STOP is excluded.

Judger	K	q	SR	SPL	OSR	NE	PL	Queries	Premature semantic STOP
Baseline	–	–	17.70	15.64	67.63	196.26	229.77	0.00	0.00
LANCE	1	1	21.47	21.47	21.47	58.89	0.02	0.77	62.07
LANCE	1	2	21.40	21.40	21.40	58.91	0.00	0.78	62.76
LANCE	1	3	21.47	21.45	21.47	58.89	0.02	0.78	62.76
LANCE	2	1	28.94	24.79	29.01	53.54	6.48	1.55	53.64
LANCE	2	2	34.77	27.47	34.77	49.29	12.94	1.53	46.23
LANCE	2	3	34.84	30.34	41.22	45.78	19.38	1.51	39.78
LANCE	3	1	34.91	27.90	34.91	49.19	12.91	2.28	46.23
LANCE	3	2	33.47	29.73	45.06	43.67	25.41	2.23	35.60
LANCE	3	3	28.94	25.20	50.00	41.92	37.13	2.17	30.66
Qwen	1	1	24.69	23.74	28.46	62.23	20.28	3.15	52.40
Qwen	1	2	25.10	23.84	29.63	62.14	22.15	2.07	50.82
Qwen	1	3	25.86	25.00	31.82	64.36	27.36	1.81	48.35
Qwen	2	1	32.24	28.06	37.52	60.86	31.28	4.46	42.59
Qwen	2	2	35.53	29.69	43.21	58.02	39.19	3.02	35.60
Qwen	2	3	33.33	28.71	47.94	58.87	49.24	2.63	31.14
Qwen	3	1	34.98	29.35	42.94	60.12	40.79	5.56	36.01
Qwen	3	2	31.76	27.86	51.65	58.92	55.29	3.96	26.68
Qwen	3	3	26.41	23.45	55.62	66.76	74.00	3.56	22.36
SenseNova	1	1	23.11	22.53	26.06	61.05	14.26	2.43	55.01
SenseNova	1	2	24.21	23.04	27.98	61.81	18.38	1.85	52.74
SenseNova	1	3	24.14	23.35	29.08	61.58	19.80	1.53	51.99
SenseNova	2	1	31.62	27.15	35.53	58.10	24.13	3.62	44.58
SenseNova	2	2	34.57	28.39	40.81	58.38	36.03	2.90	38.68
SenseNova	2	3	32.72	28.51	46.23	58.95	45.55	2.49	32.85
SenseNova	3	1	34.77	28.85	40.81	59.20	36.35	4.97	38.13
SenseNova	3	2	30.59	26.85	49.18	58.78	52.11	3.78	29.49
SenseNova	3	3	25.65	22.57	54.39	65.44	70.67	3.43	23.80

Table 11: UE486, 104 inside routes: means of three repetitions. SR, SPL and OSR are percentages. Navigation error (NE) and executed path length (PL) are measured in meters. Queries reports the mean number of logical queries per episode. The hyperparameters K and q do not apply to Baseline. Premature semantic STOP is the percentage of episodes terminated by semantic STOP before ever entering the goal region; native STOP is excluded.

Judger	K	q	SR	SPL	OSR	NE	PL	Queries	Premature semantic STOP
Baseline	–	–	40.38	33.36	100.00	150.88	156.36	0.00	0.00
LANCE	1	1	100.00	100.00	100.00	8.90	0.00	0.70	0.00
LANCE	1	2	100.00	100.00	100.00	8.90	0.00	0.69	0.00

Continued

Judger	K	q	SR	SPL	OSR	NE	PL	Queries	Premature semantic STOP
LANCE	1	3	100.00	100.00	100.00	8.90	0.00	0.69	0.00
LANCE	2	1	100.00	80.60	100.00	6.85	5.93	1.42	0.00
LANCE	2	2	100.00	66.00	100.00	9.44	12.15	1.40	0.00
LANCE	2	3	70.83	52.32	100.00	13.94	17.29	1.32	0.00
LANCE	3	1	100.00	67.38	100.00	9.20	11.86	2.03	0.00
LANCE	3	2	46.15	38.65	100.00	19.13	22.62	1.91	0.00
LANCE	3	3	42.95	36.79	100.00	27.20	31.02	1.77	0.00
Qwen	1	1	93.27	89.66	100.00	14.45	7.92	1.58	0.00
Qwen	1	2	92.95	88.21	100.00	17.17	10.77	1.25	0.00
Qwen	1	3	88.14	85.21	100.00	21.42	15.03	1.21	0.00
Qwen	2	1	90.06	71.89	100.00	21.10	21.24	3.05	0.00
Qwen	2	2	84.62	60.64	100.00	24.38	27.50	2.21	0.00
Qwen	2	3	66.35	49.99	100.00	34.14	38.01	2.01	0.00
Qwen	3	1	86.22	61.78	100.00	30.28	33.20	4.38	0.00
Qwen	3	2	46.79	38.48	100.00	40.84	44.71	3.09	0.00
Qwen	3	3	43.27	37.30	100.00	51.19	54.94	2.59	0.00
SenseNova	1	1	96.15	93.60	100.00	11.59	4.62	1.21	0.00
SenseNova	1	2	95.51	90.47	100.00	12.25	5.81	1.02	0.00
SenseNova	1	3	91.35	88.80	100.00	13.72	7.12	0.95	0.00
SenseNova	2	1	94.87	74.56	100.00	11.45	11.29	1.96	0.00
SenseNova	2	2	87.82	61.88	100.00	16.93	19.84	1.78	0.00
SenseNova	2	3	65.38	49.52	100.00	26.81	30.31	1.77	0.00
SenseNova	3	1	89.42	63.70	100.00	18.00	20.81	2.99	0.00
SenseNova	3	2	48.40	39.58	100.00	34.39	38.02	2.72	0.00
SenseNova	3	3	41.67	35.47	100.00	47.27	51.33	2.50	0.00

Table 12: UE486, 382 outside routes: means of three repetitions. SR, SPL and OSR are percentages. Navigation error (NE) and executed path length (PL) are measured in meters. Queries reports the mean number of logical queries per episode. The hyperparameters K and q do not apply to Baseline. Premature semantic STOP is the percentage of episodes terminated by semantic STOP before ever entering the goal region; native STOP is excluded.

Judger	K	q	SR	SPL	OSR	NE	PL	Queries	Premature semantic STOP
Baseline	–	–	11.52	10.82	58.81	208.61	249.76	0.00	0.00
LANCE	1	1	0.09	0.09	0.09	72.50	0.02	0.79	78.97
LANCE	1	2	0.00	0.00	0.00	72.52	0.00	0.80	79.84
LANCE	1	3	0.09	0.07	0.09	72.50	0.03	0.80	79.84
LANCE	2	1	9.60	9.60	9.69	66.25	6.62	1.59	68.24
LANCE	2	2	17.02	16.98	17.02	60.14	13.16	1.57	58.81
LANCE	2	3	25.04	24.36	25.22	54.45	19.95	1.57	50.61
LANCE	3	1	17.19	17.15	17.19	60.08	13.20	2.35	58.81
LANCE	3	2	30.02	27.30	30.10	50.35	26.18	2.31	45.29
LANCE	3	3	25.13	22.05	36.39	45.92	38.79	2.28	39.01
Qwen	1	1	6.02	5.79	8.99	75.23	23.64	3.58	66.67
Qwen	1	2	6.63	6.31	10.47	74.39	25.25	2.29	64.66
Qwen	1	3	8.90	8.60	13.26	76.05	30.72	1.97	61.52
Qwen	2	1	16.49	16.12	20.51	71.68	34.01	4.85	54.19
Qwen	2	2	22.16	21.27	27.75	67.18	42.37	3.24	45.29
Qwen	2	3	24.35	22.92	33.77	65.60	52.29	2.80	39.62
Qwen	3	1	21.03	20.52	27.40	68.24	42.86	5.88	45.81
Qwen	3	2	27.66	24.97	38.48	63.85	58.18	4.20	33.94
Qwen	3	3	21.82	19.68	43.54	71.01	79.19	3.82	28.45
SenseNova	1	1	3.23	3.18	5.93	74.51	16.88	2.76	69.98
SenseNova	1	2	4.80	4.68	8.38	75.30	21.80	2.08	67.10
SenseNova	1	3	5.85	5.53	9.77	74.61	23.25	1.69	66.14
SenseNova	2	1	14.40	14.25	17.98	70.80	27.62	4.07	56.72

Continued

Judger	K	q	SR	SPL	OSR	NE	PL	Queries	Premature semantic STOP
SenseNova	2	2	20.07	19.27	24.69	69.67	40.45	3.21	49.21
SenseNova	2	3	23.82	22.79	31.59	67.70	49.70	2.68	41.80
SenseNova	3	1	19.90	19.36	24.69	70.41	40.59	5.50	48.52
SenseNova	3	2	25.74	23.38	35.34	65.42	55.94	4.07	37.52
SenseNova	3	3	21.29	19.06	41.97	70.38	75.94	3.68	30.28

Table 13: UE586, all 586 routes: means of three repetitions. SR, SPL and OSR are percentages. Navigation error (NE) and executed path length (PL) are measured in meters. Queries reports the mean number of logical queries per episode. The hyperparameters K and q do not apply to Baseline. Premature semantic STOP is the percentage of episodes terminated by semantic STOP before ever entering the goal region; native STOP is excluded.

Judger	K	q	SR	SPL	OSR	NE	PL	Queries	Premature semantic STOP
Baseline	–	–	18.43	16.43	68.43	194.43	230.07	0.00	0.00
LANCE	1	1	22.92	22.92	22.92	58.15	0.02	0.78	61.43
LANCE	1	2	22.87	22.87	22.87	58.16	0.00	0.78	62.00
LANCE	1	3	22.92	22.91	22.92	58.15	0.02	0.78	61.95
LANCE	2	1	29.58	25.10	29.64	52.80	6.52	1.56	53.47
LANCE	2	2	35.84	27.97	35.84	48.65	13.02	1.54	45.79
LANCE	2	3	35.15	30.58	42.26	45.33	19.43	1.52	39.19
LANCE	3	1	36.06	28.43	36.06	48.58	12.99	2.29	45.51
LANCE	3	2	33.45	29.98	46.13	43.22	25.52	2.23	35.15
LANCE	3	3	29.81	25.98	51.37	41.58	37.14	2.17	29.41
Qwen	1	1	25.77	24.71	29.29	61.89	20.65	3.18	52.33
Qwen	1	2	26.22	24.72	30.55	61.85	23.03	2.11	50.68
Qwen	1	3	27.13	26.12	32.71	63.08	27.36	1.81	48.29
Qwen	2	1	32.65	27.85	37.66	60.19	31.57	4.48	43.06
Qwen	2	2	36.46	30.16	44.25	57.38	39.42	3.04	35.27
Qwen	2	3	33.56	29.33	49.15	58.71	50.27	2.67	30.38
Qwen	3	1	35.61	29.39	43.52	59.75	41.15	5.63	36.01
Qwen	3	2	32.54	28.93	52.73	58.72	56.22	4.00	26.11
Qwen	3	3	27.36	24.20	56.83	67.65	75.87	3.63	21.73
SenseNova	1	1	24.57	23.93	28.04	61.19	16.33	2.67	53.98
SenseNova	1	2	25.77	24.42	29.75	61.59	20.12	1.96	52.05
SenseNova	1	3	25.09	24.32	30.77	61.99	22.21	1.63	51.37
SenseNova	2	1	32.20	27.35	36.63	59.32	27.10	3.96	44.37
SenseNova	2	2	35.32	28.60	42.21	59.12	38.76	3.05	37.77
SenseNova	2	3	33.33	29.20	47.72	60.10	48.69	2.61	31.91
SenseNova	3	1	35.15	28.73	41.98	60.68	39.48	5.35	37.43
SenseNova	3	2	30.77	27.36	50.51	60.17	54.96	3.94	28.50
SenseNova	3	3	26.51	23.34	55.97	67.52	73.81	3.55	22.35

Table 14: UE586, 134 inside routes: means of three repetitions. SR, SPL and OSR are percentages. Navigation error (NE) and executed path length (PL) are measured in meters. Queries reports the mean number of logical queries per episode. The hyperparameters K and q do not apply to Baseline. Premature semantic STOP is the percentage of episodes terminated by semantic STOP before ever entering the goal region; native STOP is excluded.

Judger	K	q	SR	SPL	OSR	NE	PL	Queries	Premature semantic STOP
Baseline	–	–	40.55	34.06	100.00	148.13	157.68	0.00	0.00
LANCE	1	1	100.00	100.00	100.00	9.05	0.00	0.71	0.00
LANCE	1	2	100.00	100.00	100.00	9.05	0.00	0.70	0.00

Continued

Judger	K	q	SR	SPL	OSR	NE	PL	Queries	Premature semantic STOP
LANCE	1	3	100.00	100.00	100.00	9.05	0.00	0.70	0.00
LANCE	2	1	100.00	80.40	100.00	6.80	6.18	1.45	0.00
LANCE	2	2	100.00	65.73	100.00	9.49	12.42	1.43	0.00
LANCE	2	3	69.40	51.40	100.00	14.17	17.79	1.34	0.00
LANCE	3	1	100.00	66.72	100.00	9.30	12.22	2.08	0.00
LANCE	3	2	44.78	38.12	100.00	19.36	23.14	1.96	0.00
LANCE	3	3	42.29	36.59	100.00	27.50	31.50	1.82	0.00
Qwen	1	1	93.78	89.94	100.00	17.39	11.24	1.96	0.00
Qwen	1	2	93.03	87.65	100.00	20.02	14.09	1.46	0.00
Qwen	1	3	88.81	85.46	100.00	22.26	16.13	1.28	0.00
Qwen	2	1	90.55	70.88	100.00	21.46	22.01	3.15	0.00
Qwen	2	2	84.08	59.18	100.00	26.90	30.23	2.38	0.00
Qwen	2	3	60.95	47.10	100.00	38.38	42.63	2.21	0.00
Qwen	3	1	85.07	59.56	100.00	31.66	34.90	4.57	0.00
Qwen	3	2	45.77	38.58	100.00	42.53	46.46	3.20	0.00
Qwen	3	3	42.54	36.76	100.00	58.63	63.06	2.93	0.00
SenseNova	1	1	95.52	92.97	100.00	15.53	8.79	1.67	0.00
SenseNova	1	2	95.52	90.04	100.00	15.98	9.75	1.25	0.00
SenseNova	1	3	91.29	88.83	100.00	17.26	10.84	1.10	0.00
SenseNova	2	1	95.27	74.64	100.00	14.78	14.84	2.38	0.00
SenseNova	2	2	88.06	61.07	100.00	19.27	22.34	1.94	0.00
SenseNova	2	3	63.43	48.67	100.00	30.03	33.71	1.90	0.00
SenseNova	3	1	88.56	62.13	100.00	21.25	24.28	3.40	0.00
SenseNova	3	2	46.27	38.91	100.00	37.33	41.26	2.92	0.00
SenseNova	3	3	41.04	35.38	100.00	49.88	54.12	2.63	0.00

Table 15: UE586, 452 outside routes: means of three repetitions. SR, SPL and OSR are percentages. Navigation error (NE) and executed path length (PL) are measured in meters. Queries reports the mean number of logical queries per episode. The hyperparameters K and q do not apply to Baseline. Premature semantic STOP is the percentage of episodes terminated by semantic STOP before ever entering the goal region; native STOP is excluded.

Judger	K	q	SR	SPL	OSR	NE	PL	Queries	Premature semantic STOP
Baseline	–	–	11.87	11.20	59.07	208.15	251.53	0.00	0.00
LANCE	1	1	0.07	0.07	0.07	72.70	0.02	0.80	79.65
LANCE	1	2	0.00	0.00	0.00	72.72	0.00	0.80	80.38
LANCE	1	3	0.07	0.06	0.07	72.71	0.03	0.81	80.31
LANCE	2	1	8.70	8.70	8.78	66.44	6.62	1.60	69.32
LANCE	2	2	16.81	16.78	16.81	60.27	13.19	1.57	59.37
LANCE	2	3	25.00	24.40	25.15	54.57	19.91	1.57	50.81
LANCE	3	1	17.11	17.08	17.11	60.22	13.21	2.36	59.00
LANCE	3	2	30.09	27.57	30.16	50.30	26.22	2.32	45.58
LANCE	3	3	26.11	22.84	36.95	45.76	38.81	2.28	38.13
Qwen	1	1	5.60	5.38	8.33	75.09	23.44	3.54	67.85
Qwen	1	2	6.42	6.07	9.96	74.25	25.67	2.31	65.71
Qwen	1	3	8.85	8.53	12.76	75.18	30.69	1.97	62.61
Qwen	2	1	15.49	15.09	19.17	71.68	34.40	4.88	55.83
Qwen	2	2	22.35	21.55	27.73	66.42	42.14	3.23	45.72
Qwen	2	3	25.44	24.06	34.07	64.74	52.54	2.81	39.38
Qwen	3	1	20.94	20.45	26.77	68.08	43.00	5.95	46.68
Qwen	3	2	28.61	26.07	38.72	63.52	59.11	4.24	33.85
Qwen	3	3	22.86	20.48	44.03	70.32	79.66	3.84	28.17
SenseNova	1	1	3.54	3.46	6.71	74.73	18.57	2.96	69.99
SenseNova	1	2	5.09	4.96	8.92	75.11	23.19	2.17	67.48
SenseNova	1	3	5.46	5.19	10.25	75.25	25.58	1.78	66.59
SenseNova	2	1	13.50	13.34	17.85	72.52	30.73	4.43	57.52

Continued

Judger	K	q	SR	SPL	OSR	NE	PL	Queries	Premature semantic STOP
SenseNova	2	2	19.69	18.97	25.07	70.93	43.63	3.38	48.97
SenseNova	2	3	24.41	23.43	32.23	69.01	53.14	2.82	41.37
SenseNova	3	1	19.32	18.83	24.78	72.37	43.98	5.93	48.53
SenseNova	3	2	26.18	23.94	35.84	66.94	59.03	4.24	36.95
SenseNova	3	3	22.20	19.77	42.92	72.75	79.65	3.82	28.98

Table 16: AirSim180, all 180 routes: means of three repetitions. SR, SPL and OSR are percentages. Navigation error (NE) and executed path length (PL) are measured in meters. Queries reports the mean number of logical queries per episode. The hyperparameters K and q do not apply to Baseline. Premature semantic STOP is the percentage of episodes terminated by semantic STOP before ever entering the goal region; native STOP is excluded.

Judger	K	q	SR	SPL	OSR	NE	PL	Queries	Premature semantic STOP
Baseline	–	–	23.89	22.49	51.11	132.98	161.99	0.00	0.00
LANCE	1	1	7.22	7.22	7.22	78.93	0.00	0.85	79.63
LANCE	1	2	7.22	7.22	7.22	78.93	0.00	0.83	77.59
LANCE	1	3	7.22	7.22	7.22	78.93	0.00	0.83	78.89
LANCE	2	1	11.67	11.26	11.67	72.74	6.70	1.64	72.41
LANCE	2	2	16.48	14.94	16.48	67.10	13.29	1.61	65.19
LANCE	2	3	20.93	18.94	21.11	61.90	19.52	1.58	59.81
LANCE	3	1	17.22	15.85	17.22	67.06	13.27	2.42	64.63
LANCE	3	2	22.59	21.22	24.81	57.90	25.12	2.30	52.59
LANCE	3	3	29.07	27.60	33.70	51.90	34.94	2.16	42.22
Qwen	1	1	9.81	9.35	12.04	78.65	24.92	4.35	64.81
Qwen	1	2	12.41	12.20	15.56	75.09	22.64	2.51	61.67
Qwen	1	3	12.04	11.71	14.63	78.75	27.94	2.00	59.81
Qwen	2	1	14.26	13.36	18.33	77.70	39.41	6.13	55.00
Qwen	2	2	20.74	18.96	26.11	73.00	47.32	3.89	45.37
Qwen	2	3	25.00	22.25	31.48	70.92	54.43	3.00	37.22
Qwen	3	1	21.67	19.62	26.30	74.01	48.72	7.39	45.00
Qwen	3	2	25.19	23.45	34.26	75.70	64.62	4.90	32.22
Qwen	3	3	26.48	24.21	40.19	75.66	78.31	3.94	27.04
SenseNova	1	1	8.89	8.75	12.78	79.92	23.43	3.65	65.93
SenseNova	1	2	10.56	10.29	13.70	83.04	27.47	2.52	64.44
SenseNova	1	3	11.67	11.38	15.74	80.92	29.18	2.01	63.89
SenseNova	2	1	15.93	15.06	21.11	79.59	41.01	5.92	53.70
SenseNova	2	2	22.59	20.87	27.96	75.40	47.32	3.68	44.44
SenseNova	2	3	24.63	22.30	31.67	74.92	59.22	3.14	38.52
SenseNova	3	1	22.04	20.24	27.22	76.19	50.04	7.08	44.44
SenseNova	3	2	25.93	24.08	35.19	73.73	61.69	4.42	33.89
SenseNova	3	3	26.67	24.45	40.56	79.63	83.56	4.18	25.93

Table 17: AirSim180, 13 inside routes: means of three repetitions. SR, SPL and OSR are percentages. Navigation error (NE) and executed path length (PL) are measured in meters. Queries reports the mean number of logical queries per episode. The hyperparameters K and q do not apply to Baseline. Premature semantic STOP is the percentage of episodes terminated by semantic STOP before ever entering the goal region; native STOP is excluded.

Judger	K	q	SR	SPL	OSR	NE	PL	Queries	Premature semantic STOP
Baseline	–	–	53.85	49.03	100.00	52.48	67.38	0.00	0.00
LANCE	1	1	100.00	100.00	100.00	10.58	0.00	0.69	0.00
LANCE	1	2	100.00	100.00	100.00	10.58	0.00	0.69	0.00

Continued

Judger	K	q	SR	SPL	OSR	NE	PL	Queries	Premature semantic STOP
LANCE	1	3	100.00	100.00	100.00	10.58	0.00	0.62	0.00
LANCE	2	1	100.00	94.34	100.00	7.16	5.08	1.21	0.00
LANCE	2	2	100.00	78.63	100.00	7.45	11.31	1.26	0.00
LANCE	2	3	97.44	70.37	100.00	10.43	15.15	1.10	0.00
LANCE	3	1	100.00	81.03	100.00	6.75	11.08	1.82	0.00
LANCE	3	2	69.23	58.62	100.00	13.46	18.23	1.56	0.00
LANCE	3	3	56.41	54.80	100.00	19.09	23.00	1.33	0.00
Qwen	1	1	100.00	94.87	100.00	11.04	1.38	0.82	0.00
Qwen	1	2	97.44	97.13	100.00	11.07	0.92	0.77	0.00
Qwen	1	3	97.44	95.30	100.00	11.50	1.23	0.69	0.00
Qwen	2	1	97.44	87.27	100.00	8.40	6.77	1.49	0.00
Qwen	2	2	97.44	74.62	100.00	8.29	12.92	1.36	0.00
Qwen	2	3	94.87	67.04	100.00	11.43	15.54	1.13	0.00
Qwen	3	1	100.00	76.08	100.00	7.98	12.00	1.92	0.00
Qwen	3	2	64.10	53.40	100.00	14.60	19.62	1.67	0.00
Qwen	3	3	61.54	54.48	100.00	23.88	29.69	1.69	0.00
SenseNova	1	1	94.87	94.87	100.00	11.19	1.85	0.87	0.00
SenseNova	1	2	100.00	97.86	100.00	10.66	0.69	0.67	0.00
SenseNova	1	3	97.44	97.13	100.00	11.07	0.92	0.67	0.00
SenseNova	2	1	97.44	88.02	100.00	7.91	6.54	1.41	0.00
SenseNova	2	2	100.00	80.73	100.00	6.93	11.08	1.21	0.00
SenseNova	2	3	92.31	67.26	100.00	11.87	16.00	1.21	0.00
SenseNova	3	1	97.44	77.67	100.00	7.94	11.31	1.92	0.00
SenseNova	3	2	66.67	56.05	100.00	13.95	18.15	1.59	0.00
SenseNova	3	3	61.54	55.76	100.00	22.69	25.77	1.41	0.00

Table 18: AirSim180, 167 outside routes: means of three repetitions. SR, SPL and OSR are percentages. Navigation error (NE) and executed path length (PL) are measured in meters. Queries reports the mean number of logical queries per episode. The hyperparameters K and q do not apply to Baseline. Premature semantic STOP is the percentage of episodes terminated by semantic STOP before ever entering the goal region; native STOP is excluded.

Judger	K	q	SR	SPL	OSR	NE	PL	Queries	Premature semantic STOP
Baseline	–	–	21.56	20.42	47.31	139.24	169.36	0.00	0.00
LANCE	1	1	0.00	0.00	0.00	84.25	0.00	0.86	85.83
LANCE	1	2	0.00	0.00	0.00	84.25	0.00	0.84	83.63
LANCE	1	3	0.00	0.00	0.00	84.25	0.00	0.85	85.03
LANCE	2	1	4.79	4.79	4.79	77.85	6.83	1.68	78.04
LANCE	2	2	9.98	9.98	9.98	71.74	13.45	1.64	70.26
LANCE	2	3	14.97	14.93	14.97	65.91	19.86	1.62	64.47
LANCE	3	1	10.78	10.78	10.78	71.75	13.44	2.47	69.66
LANCE	3	2	18.96	18.31	18.96	61.36	25.66	2.36	56.69
LANCE	3	3	26.95	25.48	28.54	54.45	35.87	2.22	45.51
Qwen	1	1	2.79	2.69	5.19	83.91	26.75	4.62	69.86
Qwen	1	2	5.79	5.59	8.98	80.08	24.33	2.64	66.47
Qwen	1	3	5.39	5.21	7.98	83.99	30.02	2.10	64.47
Qwen	2	1	7.78	7.61	11.98	83.10	41.95	6.49	59.28
Qwen	2	2	14.77	14.63	20.36	78.04	49.99	4.09	48.90
Qwen	2	3	19.56	18.77	26.15	75.55	57.46	3.15	40.12
Qwen	3	1	15.57	15.23	20.56	79.15	51.58	7.82	48.50
Qwen	3	2	22.16	21.11	29.14	80.46	68.12	5.15	34.73
Qwen	3	3	23.75	21.85	35.53	79.69	82.10	4.11	29.14
SenseNova	1	1	2.20	2.05	5.99	85.27	25.11	3.87	71.06
SenseNova	1	2	3.59	3.47	6.99	88.68	29.56	2.67	69.46
SenseNova	1	3	4.99	4.71	9.18	86.36	31.38	2.11	68.86
SenseNova	2	1	9.58	9.38	14.97	85.17	43.69	6.27	57.88

Continued

Judger	K	q	SR	SPL	OSR	NE	PL	Queries	Premature semantic STOP
SenseNova	2	2	16.57	16.21	22.36	80.73	50.14	3.87	47.90
SenseNova	2	3	19.36	18.81	26.35	79.83	62.59	3.29	41.52
SenseNova	3	1	16.17	15.77	21.56	81.50	53.05	7.49	47.90
SenseNova	3	2	22.75	21.60	30.14	78.38	65.08	4.64	36.53
SenseNova	3	3	23.95	22.01	35.93	84.06	88.05	4.40	27.94

D SELECTED-CONFIGURATION RESULTS

The following tables use the UE100-selected pair for UE486 and the single fixed $K = 3, q = 3$ for all AirSim180 routes and judgers. Means, per-repetition outcomes, and paired effects use the same routes and preprocessing as the complete (K, q) hyperparameter grid evaluation. In these tables, TRACE (judger) denotes the selected configuration, and TRACE (judger), $K = 1$ denotes history-free mode with the same judger and query interval q as the selected temporal-evidence mode configuration ($K > 1$).

Table 19: UE100 development scores, equally weighted over the three VLMs and three repetitions. Selection maximizes SR, choosing $K = 2, q = 2$. Queries reports the mean number of logical queries per episode.

K	q	SR	SPL	Queries
2	1	34.11	27.25	3.92
2	2	40.33	30.82	2.82
2	3	35.89	32.21	2.54
3	1	39.11	29.58	5.19
3	2	33.78	31.74	3.72
3	3	32.22	28.23	3.44

Table 20: AirSim180 fold-specific configuration choices from five-scene development data, equally weighting Qwen and SenseNova. The table reports fold sizes and selected pairs, not aggregate out-of-fold performance. The common-configuration results use the modal pair $K = 3, q = 3$ on all six scenes.

Held-out scene	Development	Evaluation	K	q
16	150	30	3	3
18	150	30	3	3
23	150	30	3	3
26	150	30	3	2
GZ	150	30	3	3
SH	150	30	3	2

Table 21: UE486, all 486 routes: Means for Baseline, TRACE at the selected configuration, and matched history-free mode ($K = 1$). SR, SPL, and OSR are percentages. Navigation error (NE) and executed path length (PL) are measured in meters. Queries reports the mean number of logical queries per episode. Premature semantic STOP is the percentage of episodes terminated by semantic STOP before ever entering the goal region; native STOP is excluded.

Configuration	SR	SPL	OSR	NE	PL	Queries	Premature semantic STOP
Baseline	17.70	15.64	67.63	196.26	229.77	0.00	0.00
TRACE (LANCE), $K = 1$	21.40	21.40	21.40	58.91	0.00	0.78	62.76
TRACE (LANCE)	34.77	27.47	34.77	49.29	12.94	1.53	46.23

Continued

Configuration	SR	SPL	OSR	NE	PL	Queries	Premature semantic STOP
TRACE (Qwen), $K = 1$	25.10	23.84	29.63	62.14	22.15	2.07	50.82
TRACE (Qwen)	35.53	29.69	43.21	58.02	39.19	3.02	35.60
TRACE (SenseNova), $K = 1$	24.21	23.04	27.98	61.81	18.38	1.85	52.74
TRACE (SenseNova)	34.57	28.39	40.81	58.38	36.03	2.90	38.68

Table 22: UE486, all 486 routes: individual repetitions (percent).

Configuration	SR1	SR2	SR3	SPL1	SPL2	SPL3
Baseline	17.49	17.08	18.52	15.65	15.07	16.21
TRACE (LANCE), $K = 1$	21.40	21.40	21.40	21.40	21.40	21.40
TRACE (LANCE)	34.77	34.77	34.77	27.94	27.36	27.11
TRACE (Qwen), $K = 1$	25.93	24.69	24.69	24.26	23.61	23.64
TRACE (Qwen)	35.80	35.60	35.19	29.48	29.74	29.87
TRACE (SenseNova), $K = 1$	23.66	24.28	24.69	22.59	22.76	23.76
TRACE (SenseNova)	34.98	33.95	34.77	28.78	27.67	28.72

Table 23: UE486, all 486 routes: paired changes and outcome-transition counts by repetition. $\mathrm{F} \to \mathrm{S}$ and $\mathrm{S} \to \mathrm{F}$ denote failure-to-success and success-to-failure transitions from the reference to the method in each contrast. ΔSR and ΔSPL are in percentage points.

Contrast	ΔSR	ΔSPL	$\mathrm{F} \to \mathrm{S}$ (repetitions 1–3)	$\mathrm{S} \to \mathrm{F}$ (repetitions 1–3)
TRACE (LANCE) vs Baseline	17.08	11.83	111/112/109	27/26/30
TRACE (LANCE) vs $K = 1$	13.37	6.07	65/65/65	0/0/0
TRACE (Qwen) vs Baseline	17.83	14.05	109/111/103	20/21/22
TRACE (Qwen) vs $K = 1$	10.43	5.86	63/70/67	15/17/16
TRACE (SenseNova) vs Baseline	16.87	12.75	109/104/105	24/22/26
TRACE (SenseNova) vs $K = 1$	10.36	5.35	65/58/61	10/11/12

Table 24: UE486, all 486 routes: behavioral differences. Differences in OSR and Premature semantic STOP are in percentage points; Navigation error (NE) and executed path length (PL) differences are in meters. Query differences are differences in mean logical queries per episode. Premature semantic STOP is the percentage of episodes terminated by semantic STOP before ever entering the goal region; native STOP is excluded.

Contrast	Metric	Difference
TRACE (LANCE) vs Baseline	OSR	-32.85
TRACE (LANCE) vs Baseline	NE	-146.97
TRACE (LANCE) vs Baseline	PL	-216.83
TRACE (LANCE) vs Baseline	Queries	1.53
TRACE (LANCE) vs Baseline	Premature semantic STOP	46.23
TRACE (LANCE) vs $K = 1$	OSR	13.37
TRACE (LANCE) vs $K = 1$	NE	-9.62
TRACE (LANCE) vs $K = 1$	PL	12.94
TRACE (LANCE) vs $K = 1$	Queries	0.76
TRACE (LANCE) vs $K = 1$	Premature semantic STOP	-16.53
TRACE (Qwen) vs Baseline	OSR	-24.42
TRACE (Qwen) vs Baseline	NE	-138.23
TRACE (Qwen) vs Baseline	PL	-190.58
TRACE (Qwen) vs Baseline	Queries	3.02
TRACE (Qwen) vs Baseline	Premature semantic STOP	35.60

Continued

Contrast	Metric	Difference
TRACE (Qwen) vs $K = 1$	OSR	13.58
TRACE (Qwen) vs $K = 1$	NE	-4.12
TRACE (Qwen) vs $K = 1$	PL	17.04
TRACE (Qwen) vs $K = 1$	Queries	0.96
TRACE (Qwen) vs $K = 1$	Premature semantic STOP	-15.23
TRACE (SenseNova) vs Baseline	OSR	-26.82
TRACE (SenseNova) vs Baseline	NE	-137.87
TRACE (SenseNova) vs Baseline	PL	-193.74
TRACE (SenseNova) vs Baseline	Queries	2.90
TRACE (SenseNova) vs Baseline	Premature semantic STOP	38.68
TRACE (SenseNova) vs $K = 1$	OSR	12.83
TRACE (SenseNova) vs $K = 1$	NE	-3.42
TRACE (SenseNova) vs $K = 1$	PL	17.66
TRACE (SenseNova) vs $K = 1$	Queries	1.05
TRACE (SenseNova) vs $K = 1$	Premature semantic STOP	-14.06

Table 25: UE486, 382 outside routes: Means for Baseline, TRACE at the selected configuration, and matched history-free mode ($K = 1$). SR, SPL, and OSR are percentages. Navigation error (NE) and executed path length (PL) are measured in meters. Queries reports the mean number of logical queries per episode. Premature semantic STOP is the percentage of episodes terminated by semantic STOP before ever entering the goal region; native STOP is excluded.

Configuration	SR	SPL	OSR	NE	PL	Queries	Premature semantic STOP
Baseline	11.52	10.82	58.81	208.61	249.76	0.00	0.00
TRACE (LANCE), $K = 1$	0.00	0.00	0.00	72.52	0.00	0.80	79.84
TRACE (LANCE)	17.02	16.98	17.02	60.14	13.16	1.57	58.81
TRACE (Qwen), $K = 1$	6.63	6.31	10.47	74.39	25.25	2.29	64.66
TRACE (Qwen)	22.16	21.27	27.75	67.18	42.37	3.24	45.29
TRACE (SenseNova), $K = 1$	4.80	4.68	8.38	75.30	21.80	2.08	67.10
TRACE (SenseNova)	20.07	19.27	24.69	69.67	40.45	3.21	49.21

Table 26: UE486, 382 outside routes: individual repetitions (percent).

Configuration	SR1	SR2	SR3	SPL1	SPL2	SPL3
Baseline	11.78	10.73	12.04	11.04	10.16	11.25
TRACE (LANCE), $K = 1$	0.00	0.00	0.00	0.00	0.00	0.00
TRACE (LANCE)	17.02	17.02	17.02	16.90	17.02	17.02
TRACE (Qwen), $K = 1$	7.33	5.76	6.81	6.93	5.51	6.49
TRACE (Qwen)	21.73	22.77	21.99	20.71	21.82	21.28
TRACE (SenseNova), $K = 1$	4.19	4.97	5.24	4.15	4.89	4.99
TRACE (SenseNova)	20.68	19.37	20.16	19.77	18.55	19.49

Table 27: UE486, 382 outside routes: paired changes and outcome-transition counts by repetition. F→S and S→F denote failure-to-success and success-to-failure transitions from the reference to the method in each contrast. ΔSR and ΔSPL are in percentage points.

Contrast	ΔSR	ΔSPL	F→S (repetitions 1–3)	S→F (repetitions 1–3)
TRACE (LANCE) vs Baseline	5.50	6.16	47/50/49	27/26/30
TRACE (LANCE) vs $K = 1$	17.02	16.98	65/65/65	0/0/0
TRACE (Qwen) vs Baseline	10.65	10.45	58/64/59	20/18/21

Continued

Contrast	ΔSR	ΔSPL	F→S (repetitions 1–3)	S→F (repetitions 1–3)
TRACE (Qwen) vs $K = 1$	15.53	14.96	62/70/65	7/5/7
TRACE (SenseNova) vs Baseline	8.55	8.45	57/55/57	23/22/26
TRACE (SenseNova) vs $K = 1$	15.27	14.59	64/57/60	1/2/3

Table 28: UE486, 382 outside routes: behavioral differences. Differences in OSR and Premature semantic STOP are in percentage points; Navigation error (NE) and executed path length (PL) differences are in meters. Query differences are differences in mean logical queries per episode. Premature semantic STOP is the percentage of episodes terminated by semantic STOP before ever entering the goal region; native STOP is excluded.

Contrast	Metric	Difference
TRACE (LANCE) vs Baseline	OSR	-41.80
TRACE (LANCE) vs Baseline	NE	-148.47
TRACE (LANCE) vs Baseline	PL	-236.60
TRACE (LANCE) vs Baseline	Queries	1.57
TRACE (LANCE) vs Baseline	Premature semantic STOP	58.81
TRACE (LANCE) vs $K = 1$	OSR	17.02
TRACE (LANCE) vs $K = 1$	NE	-12.38
TRACE (LANCE) vs $K = 1$	PL	13.16
TRACE (LANCE) vs $K = 1$	Queries	0.77
TRACE (LANCE) vs $K = 1$	Premature semantic STOP	-21.03
TRACE (Qwen) vs Baseline	OSR	-31.06
TRACE (Qwen) vs Baseline	NE	-141.43
TRACE (Qwen) vs Baseline	PL	-207.39
TRACE (Qwen) vs Baseline	Queries	3.24
TRACE (Qwen) vs Baseline	Premature semantic STOP	45.29
TRACE (Qwen) vs $K = 1$	OSR	17.28
TRACE (Qwen) vs $K = 1$	NE	-7.20
TRACE (Qwen) vs $K = 1$	PL	17.13
TRACE (Qwen) vs $K = 1$	Queries	0.96
TRACE (Qwen) vs $K = 1$	Premature semantic STOP	-19.37
TRACE (SenseNova) vs Baseline	OSR	-34.12
TRACE (SenseNova) vs Baseline	NE	-138.94
TRACE (SenseNova) vs Baseline	PL	-209.32
TRACE (SenseNova) vs Baseline	Queries	3.21
TRACE (SenseNova) vs Baseline	Premature semantic STOP	49.21
TRACE (SenseNova) vs $K = 1$	OSR	16.32
TRACE (SenseNova) vs $K = 1$	NE	-5.63
TRACE (SenseNova) vs $K = 1$	PL	18.65
TRACE (SenseNova) vs $K = 1$	Queries	1.13
TRACE (SenseNova) vs $K = 1$	Premature semantic STOP	-17.89

Table 29: UE486, 104 inside routes: Means for Baseline, TRACE at the selected configuration, and matched history-free mode ($K = 1$). SR, SPL, and OSR are percentages. Navigation error (NE) and executed path length (PL) are measured in meters. Queries reports the mean number of logical queries per episode. Premature semantic STOP is the percentage of episodes terminated by semantic STOP before ever entering the goal region; native STOP is excluded.

Configuration	SR	SPL	OSR	NE	PL	Queries	Premature semantic STOP
Baseline	40.38	33.36	100.00	150.88	156.36	0.00	0.00
TRACE (LANCE), $K = 1$	100.00	100.00	100.00	8.90	0.00	0.69	0.00
TRACE (LANCE)	100.00	66.00	100.00	9.44	12.15	1.40	0.00

Continued

Configuration	SR	SPL	OSR	NE	PL	Queries	Premature semantic STOP
TRACE (Qwen), $K = 1$	92.95	88.21	100.00	17.17	10.77	1.25	0.00
TRACE (Qwen)	84.62	60.64	100.00	24.38	27.50	2.21	0.00
TRACE (SenseNova), $K = 1$	95.51	90.47	100.00	12.25	5.81	1.02	0.00
TRACE (SenseNova)	87.82	61.88	100.00	16.93	19.84	1.78	0.00

Table 30: UE486, 104 inside routes: individual repetitions (percent).

Configuration	SR1	SR2	SR3	SPL1	SPL2	SPL3
Baseline	38.46	40.38	42.31	32.58	33.08	34.42
TRACE (LANCE), $K = 1$	100.00	100.00	100.00	100.00	100.00	100.00
TRACE (LANCE)	100.00	100.00	100.00	68.47	65.34	64.19
TRACE (Qwen), $K = 1$	94.23	94.23	90.38	87.93	90.08	86.64
TRACE (Qwen)	87.50	82.69	83.65	61.70	58.80	61.41
TRACE (SenseNova), $K = 1$	95.19	95.19	96.15	90.32	88.38	92.70
TRACE (SenseNova)	87.50	87.50	88.46	61.88	61.16	62.61

Table 31: UE486, 104 inside routes: paired changes and outcome-transition counts by repetition. $\mathrm{F}\rightarrow\mathrm{S}$ and $\mathrm{S}\rightarrow\mathrm{F}$ denote failure-to-success and success-to-failure transitions from the reference to the method in each contrast. ΔSR and ΔSPL are in percentage points.

Contrast	ΔSR	ΔSPL	$\mathrm{F}\rightarrow\mathrm{S}$ (repetitions 1–3)	$\mathrm{S}\rightarrow\mathrm{F}$ (repetitions 1–3)
TRACE (LANCE) vs Baseline	59.62	32.64	64/62/60	0/0/0
TRACE (LANCE) vs $K = 1$	0.00	-34.00	0/0/0	0/0/0
TRACE (Qwen) vs Baseline	44.23	27.27	51/47/44	0/3/1
TRACE (Qwen) vs $K = 1$	-8.33	-27.58	1/0/2	8/12/9
TRACE (SenseNova) vs Baseline	47.44	28.52	52/49/48	1/0/0
TRACE (SenseNova) vs $K = 1$	-7.69	-28.59	1/1/1	9/9/9

Table 32: UE486, 104 inside routes: behavioral differences. Differences in OSR and Premature semantic STOP are in percentage points; Navigation error (NE) and executed path length (PL) differences are in meters. Query differences are differences in mean logical queries per episode. Premature semantic STOP is the percentage of episodes terminated by semantic STOP before ever entering the goal region; native STOP is excluded.

Contrast	Metric	Difference
TRACE (LANCE) vs Baseline	OSR	0.00
TRACE (LANCE) vs Baseline	NE	-141.44
TRACE (LANCE) vs Baseline	PL	-144.20
TRACE (LANCE) vs Baseline	Queries	1.40
TRACE (LANCE) vs Baseline	Premature semantic STOP	0.00
TRACE (LANCE) vs $K = 1$	OSR	0.00
TRACE (LANCE) vs $K = 1$	NE	0.54
TRACE (LANCE) vs $K = 1$	PL	12.15
TRACE (LANCE) vs $K = 1$	Queries	0.71
TRACE (LANCE) vs $K = 1$	Premature semantic STOP	0.00
TRACE (Qwen) vs Baseline	OSR	0.00
TRACE (Qwen) vs Baseline	NE	-126.50
TRACE (Qwen) vs Baseline	PL	-128.86
TRACE (Qwen) vs Baseline	Queries	2.21
TRACE (Qwen) vs Baseline	Premature semantic STOP	0.00

Continued

Contrast	Metric	Difference
TRACE (Qwen) vs $K = 1$	OSR	0.00
TRACE (Qwen) vs $K = 1$	NE	7.21
TRACE (Qwen) vs $K = 1$	PL	16.73
TRACE (Qwen) vs $K = 1$	Queries	0.96
TRACE (Qwen) vs $K = 1$	Premature semantic STOP	0.00
TRACE (SenseNova) vs Baseline	OSR	0.00
TRACE (SenseNova) vs Baseline	NE	-133.95
TRACE (SenseNova) vs Baseline	PL	-136.52
TRACE (SenseNova) vs Baseline	Queries	1.78
TRACE (SenseNova) vs Baseline	Premature semantic STOP	0.00
TRACE (SenseNova) vs $K = 1$	OSR	0.00
TRACE (SenseNova) vs $K = 1$	NE	4.68
TRACE (SenseNova) vs $K = 1$	PL	14.03
TRACE (SenseNova) vs $K = 1$	Queries	0.76
TRACE (SenseNova) vs $K = 1$	Premature semantic STOP	0.00

Table 33: AirSim180, all 180 routes: Means for Baseline, TRACE at the selected configuration, and matched history-free mode ($K = 1$). SR, SPL, and OSR are percentages. Navigation error (NE) and executed path length (PL) are measured in meters. Queries reports the mean number of logical queries per episode. Premature semantic STOP is the percentage of episodes terminated by semantic STOP before ever entering the goal region; native STOP is excluded.

Configuration	SR	SPL	OSR	NE	PL	Queries	Premature semantic STOP
Baseline	23.89	22.49	51.11	132.98	161.99	0.00	0.00
TRACE (LANCE), $K = 1$	7.22	7.22	7.22	78.93	0.00	0.83	78.89
TRACE (LANCE)	29.07	27.60	33.70	51.90	34.94	2.16	42.22
TRACE (Qwen), $K = 1$	12.04	11.71	14.63	78.75	27.94	2.00	59.81
TRACE (Qwen)	26.48	24.21	40.19	75.66	78.31	3.94	27.04
TRACE (SenseNova), $K = 1$	11.67	11.38	15.74	80.92	29.18	2.01	63.89
TRACE (SenseNova)	26.67	24.45	40.56	79.63	83.56	4.18	25.93

Table 34: AirSim180, all 180 routes: individual repetitions (percent).

Configuration	SR1	SR2	SR3	SPL1	SPL2	SPL3
Baseline	20.56	24.44	26.67	19.12	22.87	25.47
TRACE (LANCE), $K = 1$	7.22	7.22	7.22	7.22	7.22	7.22
TRACE (LANCE)	27.78	28.33	31.11	26.55	26.79	29.46
TRACE (Qwen), $K = 1$	12.78	10.56	12.78	12.64	10.40	12.10
TRACE (Qwen)	26.11	27.22	26.11	23.76	24.87	23.99
TRACE (SenseNova), $K = 1$	11.67	11.67	11.67	11.36	11.54	11.25
TRACE (SenseNova)	27.78	26.11	26.11	26.21	23.48	23.65

Table 35: AirSim180, all 180 routes: paired changes and outcome-transition counts by repetition. F→S and S→F denote failure-to-success and success-to-failure transitions from the reference to the method in each contrast. ΔSR and ΔSPL are in percentage points.

Contrast	ΔSR	ΔSPL	F→S (repetitions 1–3)	S→F (repetitions 1–3)
TRACE (LANCE) vs Baseline	5.19	5.11	27/24/29	14/17/21
TRACE (LANCE) vs $K = 1$	21.85	20.38	42/44/49	5/6/6
TRACE (Qwen) vs Baseline	2.59	1.72	23/20/17	13/15/18

Continued

Contrast	ΔSR	ΔSPL	F→S (repetitions 1–3)	S→F (repetitions 1–3)
TRACE (Qwen) vs $K = 1$	14.44	12.49	33/37/32	9/7/8
TRACE (SenseNova) vs Baseline	2.78	1.96	25/19/19	12/16/20
TRACE (SenseNova) vs $K = 1$	15.00	13.07	36/35/36	7/9/10

Table 36: AirSim180, all 180 routes: behavioral differences. Differences in OSR and Premature semantic STOP are in percentage points; Navigation error (NE) and executed path length (PL) differences are in meters. Query differences are differences in mean logical queries per episode. Premature semantic STOP is the percentage of episodes terminated by semantic STOP before ever entering the goal region; native STOP is excluded.

Contrast	Metric	Difference
TRACE (LANCE) vs Baseline	OSR	-17.41
TRACE (LANCE) vs Baseline	NE	-81.07
TRACE (LANCE) vs Baseline	PL	-127.06
TRACE (LANCE) vs Baseline	Queries	2.16
TRACE (LANCE) vs Baseline	Premature semantic STOP	42.22
TRACE (LANCE) vs $K = 1$	OSR	26.48
TRACE (LANCE) vs $K = 1$	NE	-27.03
TRACE (LANCE) vs $K = 1$	PL	34.94
TRACE (LANCE) vs $K = 1$	Queries	1.32
TRACE (LANCE) vs $K = 1$	Premature semantic STOP	-36.67
TRACE (Qwen) vs Baseline	OSR	-10.93
TRACE (Qwen) vs Baseline	NE	-57.31
TRACE (Qwen) vs Baseline	PL	-83.68
TRACE (Qwen) vs Baseline	Queries	3.94
TRACE (Qwen) vs Baseline	Premature semantic STOP	27.04
TRACE (Qwen) vs $K = 1$	OSR	25.56
TRACE (Qwen) vs $K = 1$	NE	-3.09
TRACE (Qwen) vs $K = 1$	PL	50.37
TRACE (Qwen) vs $K = 1$	Queries	1.94
TRACE (Qwen) vs $K = 1$	Premature semantic STOP	-32.78
TRACE (SenseNova) vs Baseline	OSR	-10.56
TRACE (SenseNova) vs Baseline	NE	-53.34
TRACE (SenseNova) vs Baseline	PL	-78.44
TRACE (SenseNova) vs Baseline	Queries	4.18
TRACE (SenseNova) vs Baseline	Premature semantic STOP	25.93
TRACE (SenseNova) vs $K = 1$	OSR	24.81
TRACE (SenseNova) vs $K = 1$	NE	-1.29
TRACE (SenseNova) vs $K = 1$	PL	54.37
TRACE (SenseNova) vs $K = 1$	Queries	2.18
TRACE (SenseNova) vs $K = 1$	Premature semantic STOP	-37.96

Table 37: AirSim180, 167 outside routes: Means for Baseline, TRACE at the selected configuration, and matched history-free mode ($K = 1$). SR, SPL, and OSR are percentages. Navigation error (NE) and executed path length (PL) are measured in meters. Queries reports the mean number of logical queries per episode. Premature semantic STOP is the percentage of episodes terminated by semantic STOP before ever entering the goal region; native STOP is excluded.

Configuration	SR	SPL	OSR	NE	PL	Queries	Premature semantic STOP
Baseline	21.56	20.42	47.31	139.24	169.36	0.00	0.00
TRACE (LANCE), $K = 1$	0.00	0.00	0.00	84.25	0.00	0.85	85.03
TRACE (LANCE)	26.95	25.48	28.54	54.45	35.87	2.22	45.51

Continued

Configuration	SR	SPL	OSR	NE	PL	Queries	Premature semantic STOP
TRACE (Qwen), $K = 1$	5.39	5.21	7.98	83.99	30.02	2.10	64.47
TRACE (Qwen)	23.75	21.85	35.53	79.69	82.10	4.11	29.14
TRACE (SenseNova), $K = 1$	4.99	4.71	9.18	86.36	31.38	2.11	68.86
TRACE (SenseNova)	23.95	22.01	35.93	84.06	88.05	4.40	27.94

Table 38: AirSim180, 167 outside routes: individual repetitions (percent).

Configuration	SR1	SR2	SR3	SPL1	SPL2	SPL3
Baseline	18.56	22.16	23.95	17.54	20.99	22.73
TRACE (LANCE), $K = 1$	0.00	0.00	0.00	0.00	0.00	0.00
TRACE (LANCE)	25.15	26.35	29.34	23.82	24.72	27.90
TRACE (Qwen), $K = 1$	5.99	4.19	5.99	5.84	4.02	5.76
TRACE (Qwen)	23.35	24.55	23.35	21.63	22.35	21.57
TRACE (SenseNova), $K = 1$	4.79	4.79	5.39	4.46	4.65	5.01
TRACE (SenseNova)	25.15	23.95	22.75	23.80	21.62	20.62

Table 39: AirSim180, 167 outside routes: paired changes and outcome-transition counts by repetition. $\mathrm{F} \rightarrow \mathrm{S}$ and $\mathrm{S} \rightarrow \mathrm{F}$ denote failure-to-success and success-to-failure transitions from the reference to the method in each contrast. ΔSR and ΔSPL are in percentage points.

Contrast	ΔSR	ΔSPL	$\mathrm{F} \rightarrow \mathrm{S}$ (repetitions 1–3)	$\mathrm{S} \rightarrow \mathrm{F}$ (repetitions 1–3)
TRACE (LANCE) vs Baseline	5.39	5.06	25/24/29	14/17/20
TRACE (LANCE) vs $K = 1$	26.95	25.48	42/44/49	0/0/0
TRACE (Qwen) vs Baseline	2.20	1.43	21/19/17	13/15/18
TRACE (Qwen) vs $K = 1$	18.36	16.65	33/36/32	4/2/3
TRACE (SenseNova) vs Baseline	2.40	1.59	23/19/17	12/16/19
TRACE (SenseNova) vs $K = 1$	18.96	17.30	36/35/35	2/3/6

Table 40: AirSim180, 167 outside routes: behavioral differences. Differences in OSR and Premature semantic STOP are in percentage points; Navigation error (NE) and executed path length (PL) differences are in meters. Query differences are differences in mean logical queries per episode. Premature semantic STOP is the percentage of episodes terminated by semantic STOP before ever entering the goal region; native STOP is excluded.

Contrast	Metric	Difference
TRACE (LANCE) vs Baseline	OSR	-18.76
TRACE (LANCE) vs Baseline	NE	-84.79
TRACE (LANCE) vs Baseline	PL	-133.49
TRACE (LANCE) vs Baseline	Queries	2.22
TRACE (LANCE) vs Baseline	Premature semantic STOP	45.51
TRACE (LANCE) vs $K = 1$	OSR	28.54
TRACE (LANCE) vs $K = 1$	NE	-29.79
TRACE (LANCE) vs $K = 1$	PL	35.87
TRACE (LANCE) vs $K = 1$	Queries	1.37
TRACE (LANCE) vs $K = 1$	Premature semantic STOP	-39.52
TRACE (Qwen) vs Baseline	OSR	-11.78
TRACE (Qwen) vs Baseline	NE	-59.55
TRACE (Qwen) vs Baseline	PL	-87.26
TRACE (Qwen) vs Baseline	Queries	4.11
TRACE (Qwen) vs Baseline	Premature semantic STOP	29.14

Continued

Contrast	Metric	Difference
TRACE (Qwen) vs $K = 1$	OSR	27.54
TRACE (Qwen) vs $K = 1$	NE	-4.29
TRACE (Qwen) vs $K = 1$	PL	52.07
TRACE (Qwen) vs $K = 1$	Queries	2.01
TRACE (Qwen) vs $K = 1$	Premature semantic STOP	-35.33
TRACE (SenseNova) vs Baseline	OSR	-11.38
TRACE (SenseNova) vs Baseline	NE	-55.18
TRACE (SenseNova) vs Baseline	PL	-81.31
TRACE (SenseNova) vs Baseline	Queries	4.40
TRACE (SenseNova) vs Baseline	Premature semantic STOP	27.94
TRACE (SenseNova) vs $K = 1$	OSR	26.75
TRACE (SenseNova) vs $K = 1$	NE	-2.29
TRACE (SenseNova) vs $K = 1$	PL	56.67
TRACE (SenseNova) vs $K = 1$	Queries	2.29
TRACE (SenseNova) vs $K = 1$	Premature semantic STOP	-40.92

Table 41: AirSim180, 13 inside routes: Means for Baseline, TRACE at the selected configuration, and matched history-free mode ($K = 1$). SR, SPL, and OSR are percentages. Navigation error (NE) and executed path length (PL) are measured in meters. Queries reports the mean number of logical queries per episode. Premature semantic STOP is the percentage of episodes terminated by semantic STOP before ever entering the goal region; native STOP is excluded.

Configuration	SR	SPL	OSR	NE	PL	Queries	Premature semantic STOP
Baseline	53.85	49.03	100.00	52.48	67.38	0.00	0.00
TRACE (LANCE), $K = 1$	100.00	100.00	100.00	10.58	0.00	0.62	0.00
TRACE (LANCE)	56.41	54.80	100.00	19.09	23.00	1.33	0.00
TRACE (Qwen), $K = 1$	97.44	95.30	100.00	11.50	1.23	0.69	0.00
TRACE (Qwen)	61.54	54.48	100.00	23.88	29.69	1.69	0.00
TRACE (SenseNova), $K = 1$	97.44	97.13	100.00	11.07	0.92	0.67	0.00
TRACE (SenseNova)	61.54	55.76	100.00	22.69	25.77	1.41	0.00

Table 42: AirSim180, 13 inside routes: individual repetitions (percent).

Configuration	SR1	SR2	SR3	SPL1	SPL2	SPL3
Baseline	46.15	53.85	61.54	39.51	46.95	60.63
TRACE (LANCE), $K = 1$	100.00	100.00	100.00	100.00	100.00	100.00
TRACE (LANCE)	61.54	53.85	53.85	61.54	53.36	49.51
TRACE (Qwen), $K = 1$	100.00	92.31	100.00	100.00	92.31	93.59
TRACE (Qwen)	61.54	61.54	61.54	51.10	57.21	55.13
TRACE (SenseNova), $K = 1$	100.00	100.00	92.31	100.00	100.00	91.40
TRACE (SenseNova)	61.54	53.85	69.23	57.21	47.44	62.64

Table 43: AirSim180, 13 inside routes: paired changes and outcome-transition counts by repetition. F→S and S→F denote failure-to-success and success-to-failure transitions from the reference to the method in each contrast. ΔSR and ΔSPL are in percentage points.

Contrast	ΔSR	ΔSPL	F→S (repetitions 1–3)	S→F (repetitions 1–3)
TRACE (LANCE) vs Baseline	2.56	5.77	2/0/0	0/0/1
TRACE (LANCE) vs $K = 1$	-43.59	-45.20	0/0/0	5/6/6
TRACE (Qwen) vs Baseline	7.69	5.45	2/1/0	0/0/0

Continued

Contrast	ΔSR	ΔSPL	F→S (repetitions 1–3)	S→F (repetitions 1–3)
TRACE (Qwen) vs $K = 1$	-35.90	-40.82	0/1/0	5/5/5
TRACE (SenseNova) vs Baseline	7.69	6.73	2/0/2	0/0/1
TRACE (SenseNova) vs $K = 1$	-35.90	-41.37	0/0/1	5/6/4

Table 44: AirSim180, 13 inside routes: behavioral differences. Differences in OSR and Premature semantic STOP are in percentage points; Navigation error (NE) and executed path length (PL) differences are in meters. Query differences are differences in mean logical queries per episode. Premature semantic STOP is the percentage of episodes terminated by semantic STOP before ever entering the goal region; native STOP is excluded.

Contrast	Metric	Difference
TRACE (LANCE) vs Baseline	OSR	0.00
TRACE (LANCE) vs Baseline	NE	-33.39
TRACE (LANCE) vs Baseline	PL	-44.38
TRACE (LANCE) vs Baseline	Queries	1.33
TRACE (LANCE) vs Baseline	Premature semantic STOP	0.00
TRACE (LANCE) vs $K = 1$	OSR	0.00
TRACE (LANCE) vs $K = 1$	NE	8.51
TRACE (LANCE) vs $K = 1$	PL	23.00
TRACE (LANCE) vs $K = 1$	Queries	0.72
TRACE (LANCE) vs $K = 1$	Premature semantic STOP	0.00
TRACE (Qwen) vs Baseline	OSR	0.00
TRACE (Qwen) vs Baseline	NE	-28.60
TRACE (Qwen) vs Baseline	PL	-37.69
TRACE (Qwen) vs Baseline	Queries	1.69
TRACE (Qwen) vs Baseline	Premature semantic STOP	0.00
TRACE (Qwen) vs $K = 1$	OSR	0.00
TRACE (Qwen) vs $K = 1$	NE	12.38
TRACE (Qwen) vs $K = 1$	PL	28.46
TRACE (Qwen) vs $K = 1$	Queries	1.00
TRACE (Qwen) vs $K = 1$	Premature semantic STOP	0.00
TRACE (SenseNova) vs Baseline	OSR	0.00
TRACE (SenseNova) vs Baseline	NE	-29.79
TRACE (SenseNova) vs Baseline	PL	-41.62
TRACE (SenseNova) vs Baseline	Queries	1.41
TRACE (SenseNova) vs Baseline	Premature semantic STOP	0.00
TRACE (SenseNova) vs $K = 1$	OSR	0.00
TRACE (SenseNova) vs $K = 1$	NE	11.62
TRACE (SenseNova) vs $K = 1$	PL	24.85
TRACE (SenseNova) vs $K = 1$	Queries	0.74
TRACE (SenseNova) vs $K = 1$	Premature semantic STOP	0.00

E PAIRED OUTCOMES AND BEHAVIORAL DIFFERENCES

The following tables report every Baseline-relative comparison and every paired $K = 2, 3$ versus $K = 1$ comparison on outside routes in UE486 and AirSim180. Baseline contrasts include all nine cells for each judger; temporal-evidence mode versus history-free mode contrasts hold judger and q fixed. SR and SPL differences and per-repetition failure-to-success and success-to-failure counts accompany OSR, premature semantic STOP rate, path-length, and query-cost differences. Both transition types can occur; their net count divided by the route count gives the SR change. All differences use the route-paired procedure in Appendix A.7.

Table 45: UE486 outside: paired differences from Baseline. ΔSR and ΔSPL are in percentage points. F → S and S → F denote failure-to-success and success-to-failure transitions from Baseline to the method. Counts are listed in repetition order (1, 2, 3).

Judger	K	q	ΔSR	ΔSPL	F → S (repetitions 1–3)	S → F (repetitions 1–3)
LANCE	1	1	-11.43	-10.73	0/0/1	45/41/46
LANCE	1	2	-11.52	-10.82	0/0/0	45/41/46
LANCE	1	3	-11.43	-10.75	1/0/0	45/41/46
LANCE	2	1	-1.92	-1.22	26/28/27	36/32/35
LANCE	2	2	5.50	6.16	47/50/49	27/26/30
LANCE	2	3	13.53	13.54	74/76/73	24/20/24
LANCE	3	1	5.67	6.33	47/51/49	27/26/29
LANCE	3	2	18.50	16.48	91/93/90	22/18/22
LANCE	3	3	13.61	11.23	68/70/65	17/13/17
Qwen	1	1	-5.50	-5.03	18/17/18	39/36/41
Qwen	1	2	-4.89	-4.51	20/18/20	37/37/40
Qwen	1	3	-2.62	-2.21	24/24/23	35/29/37
Qwen	2	1	4.97	5.31	45/48/47	28/27/28
Qwen	2	2	10.65	10.45	58/64/59	20/18/21
Qwen	2	3	12.83	12.10	65/66/65	15/14/20
Qwen	3	1	9.51	9.70	56/57/55	21/18/20
Qwen	3	2	16.14	14.15	80/77/75	15/13/19
Qwen	3	3	10.30	8.86	52/49/53	11/11/14
SenseNova	1	1	-8.29	-7.63	18/6/9	43/40/45
SenseNova	1	2	-6.72	-6.14	13/16/15	42/38/41
SenseNova	1	3	-5.67	-5.28	15/18/17	38/35/42
SenseNova	2	1	2.88	3.43	37/42/39	29/26/30
SenseNova	2	2	8.55	8.45	57/55/57	23/22/26
SenseNova	2	3	12.30	11.97	64/68/68	20/19/20
SenseNova	3	1	8.38	8.54	55/55/54	21/20/27
SenseNova	3	2	14.22	12.57	71/72/74	17/18/19
SenseNova	3	3	9.77	8.24	48/53/53	15/13/14

Table 46: UE486 outside: behavior differences from Baseline. Differences in OSR and Premature semantic STOP are in percentage points; executed path length (PL) differences are in meters. Query differences are differences in mean logical queries per episode. Premature semantic STOP is the percentage of episodes terminated by semantic STOP before ever entering the goal region; native STOP is excluded.

Judger	K	q	Metric	Difference
LANCE	1	1	OSR	-58.73
LANCE	1	1	Premature semantic STOP	78.97
LANCE	1	1	PL	-249.74
LANCE	1	1	Queries	0.79
LANCE	1	2	OSR	-58.81
LANCE	1	2	Premature semantic STOP	79.84
LANCE	1	2	PL	-249.76
LANCE	1	2	Queries	0.80
LANCE	1	3	OSR	-58.73
LANCE	1	3	Premature semantic STOP	79.84
LANCE	1	3	PL	-249.73
LANCE	1	3	Queries	0.80
LANCE	2	1	OSR	-49.13
LANCE	2	1	Premature semantic STOP	68.24
LANCE	2	1	PL	-243.14
LANCE	2	1	Queries	1.59
LANCE	2	2	OSR	-41.80
LANCE	2	2	Premature semantic STOP	58.81

Continued

Judger	K	q	Metric	Difference
LANCE	2	2	PL	-236.60
LANCE	2	2	Queries	1.57
LANCE	2	3	OSR	-33.60
LANCE	2	3	Premature semantic STOP	50.61
LANCE	2	3	PL	-229.81
LANCE	2	3	Queries	1.57
LANCE	3	1	OSR	-41.62
LANCE	3	1	Premature semantic STOP	58.81
LANCE	3	1	PL	-236.57
LANCE	3	1	Queries	2.35
LANCE	3	2	OSR	-28.71
LANCE	3	2	Premature semantic STOP	45.29
LANCE	3	2	PL	-223.59
LANCE	3	2	Queries	2.31
LANCE	3	3	OSR	-22.43
LANCE	3	3	Premature semantic STOP	39.01
LANCE	3	3	PL	-210.97
LANCE	3	3	Queries	2.28
Qwen	1	1	OSR	-49.83
Qwen	1	1	Premature semantic STOP	66.67
Qwen	1	1	PL	-226.12
Qwen	1	1	Queries	3.58
Qwen	1	2	OSR	-48.34
Qwen	1	2	Premature semantic STOP	64.66
Qwen	1	2	PL	-224.52
Qwen	1	2	Queries	2.29
Qwen	1	3	OSR	-45.55
Qwen	1	3	Premature semantic STOP	61.52
Qwen	1	3	PL	-219.04
Qwen	1	3	Queries	1.97
Qwen	2	1	OSR	-38.31
Qwen	2	1	Premature semantic STOP	54.19
Qwen	2	1	PL	-215.75
Qwen	2	1	Queries	4.85
Qwen	2	2	OSR	-31.06
Qwen	2	2	Premature semantic STOP	45.29
Qwen	2	2	PL	-207.39
Qwen	2	2	Queries	3.24
Qwen	2	3	OSR	-25.04
Qwen	2	3	Premature semantic STOP	39.62
Qwen	2	3	PL	-197.47
Qwen	2	3	Queries	2.80
Qwen	3	1	OSR	-31.41
Qwen	3	1	Premature semantic STOP	45.81
Qwen	3	1	PL	-206.91
Qwen	3	1	Queries	5.88
Qwen	3	2	OSR	-20.33
Qwen	3	2	Premature semantic STOP	33.94
Qwen	3	2	PL	-191.59
Qwen	3	2	Queries	4.20
Qwen	3	3	OSR	-15.27
Qwen	3	3	Premature semantic STOP	28.45
Qwen	3	3	PL	-170.57
Qwen	3	3	Queries	3.82
SenseNova	1	1	OSR	-52.88
SenseNova	1	1	Premature semantic STOP	69.98
SenseNova	1	1	PL	-232.88
SenseNova	1	1	Queries	2.76
SenseNova	1	2	OSR	-50.44
SenseNova	1	2	Premature semantic STOP	67.10
SenseNova	1	2	PL	-227.96
SenseNova	1	2	Queries	2.08

Continued

Judger	K	q	Metric	Difference
SenseNova	1	3	OSR	-49.04
SenseNova	1	3	Premature semantic STOP	66.14
SenseNova	1	3	PL	-226.51
SenseNova	1	3	Queries	1.69
SenseNova	2	1	OSR	-40.84
SenseNova	2	1	Premature semantic STOP	56.72
SenseNova	2	1	PL	-222.14
SenseNova	2	1	Queries	4.07
SenseNova	2	2	OSR	-34.12
SenseNova	2	2	Premature semantic STOP	49.21
SenseNova	2	2	PL	-209.32
SenseNova	2	2	Queries	3.21
SenseNova	2	3	OSR	-27.23
SenseNova	2	3	Premature semantic STOP	41.80
SenseNova	2	3	PL	-200.06
SenseNova	2	3	Queries	2.68
SenseNova	3	1	OSR	-34.12
SenseNova	3	1	Premature semantic STOP	48.52
SenseNova	3	1	PL	-209.18
SenseNova	3	1	Queries	5.50
SenseNova	3	2	OSR	-23.47
SenseNova	3	2	Premature semantic STOP	37.52
SenseNova	3	2	PL	-193.82
SenseNova	3	2	Queries	4.07
SenseNova	3	3	OSR	-16.84
SenseNova	3	3	Premature semantic STOP	30.28
SenseNova	3	3	PL	-173.82
SenseNova	3	3	Queries	3.68

Table 47: AirSim180 outside: paired differences from Baseline. ΔSR and ΔSPL are in percentage points. $\mathrm{F}\rightarrow\mathrm{S}$ and $\mathrm{S}\rightarrow\mathrm{F}$ denote failure-to-success and success-to-failure transitions from Baseline to the method. Counts are listed in repetition order (1, 2, 3).

Judger	K	q	ΔSR	ΔSPL	$\mathrm{F}\rightarrow\mathrm{S}$ (repetitions 1–3)	$\mathrm{S}\rightarrow\mathrm{F}$ (repetitions 1–3)
LANCE	1	1	-21.56	-20.42	0/0/0	31/37/40
LANCE	1	2	-21.56	-20.42	0/0/0	31/37/40
LANCE	1	3	-21.56	-20.42	0/0/0	31/37/40
LANCE	2	1	-16.77	-15.63	6/4/5	29/33/37
LANCE	2	2	-11.58	-10.44	11/7/10	25/28/33
LANCE	2	3	-6.59	-5.49	14/12/13	21/23/28
LANCE	3	1	-10.78	-9.64	12/8/10	25/27/32
LANCE	3	2	-2.59	-2.11	20/20/16	19/23/27
LANCE	3	3	5.39	5.06	25/24/29	14/17/20
Qwen	1	1	-18.76	-17.73	1/5/3	29/36/38
Qwen	1	2	-15.77	-14.83	6/6/5	29/32/35
Qwen	1	3	-16.17	-15.21	6/5/3	27/35/33
Qwen	2	1	-13.77	-12.81	10/6/8	28/30/35
Qwen	2	2	-6.79	-5.79	17/10/14	23/25/27
Qwen	2	3	-2.00	-1.65	21/13/16	19/18/23
Qwen	3	1	-5.99	-5.19	19/12/15	24/23/29
Qwen	3	2	0.60	0.69	25/20/22	17/22/25
Qwen	3	3	2.20	1.43	21/19/17	13/15/18
SenseNova	1	1	-19.36	-18.37	2/3/4	30/36/40
SenseNova	1	2	-17.96	-16.95	4/5/3	30/34/38
SenseNova	1	3	-16.57	-15.71	5/4/3	28/33/34
SenseNova	2	1	-11.98	-11.04	10/6/10	24/30/32
SenseNova	2	2	-4.99	-4.21	16/13/13	19/23/25

Continued

Judger	K	q	ΔSR	ΔSPL	F→S (repetitions 1–3)	S→F (repetitions 1–3)
SenseNova	2	3	-2.20	-1.62	22/12/16	18/20/23
SenseNova	3	1	-5.39	-4.65	12/14/16	21/22/26
SenseNova	3	2	1.20	1.18	24/19/22	17/18/24
SenseNova	3	3	2.40	1.59	23/19/17	12/16/19

Table 48: AirSim180 outside: behavior differences from Baseline. Differences in OSR and Premature semantic STOP are in percentage points; executed path length (PL) differences are in meters. Query differences are differences in mean logical queries per episode. Premature semantic STOP is the percentage of episodes terminated by semantic STOP before ever entering the goal region; native STOP is excluded.

Judger	K	q	Metric	Difference
LANCE	1	1	OSR	-47.31
LANCE	1	1	Premature semantic STOP	85.83
LANCE	1	1	PL	-169.36
LANCE	1	1	Queries	0.86
LANCE	1	2	OSR	-47.31
LANCE	1	2	Premature semantic STOP	83.63
LANCE	1	2	PL	-169.36
LANCE	1	2	Queries	0.84
LANCE	1	3	OSR	-47.31
LANCE	1	3	Premature semantic STOP	85.03
LANCE	1	3	PL	-169.36
LANCE	1	3	Queries	0.85
LANCE	2	1	OSR	-42.51
LANCE	2	1	Premature semantic STOP	78.04
LANCE	2	1	PL	-162.53
LANCE	2	1	Queries	1.68
LANCE	2	2	OSR	-37.33
LANCE	2	2	Premature semantic STOP	70.26
LANCE	2	2	PL	-155.91
LANCE	2	2	Queries	1.64
LANCE	2	3	OSR	-32.34
LANCE	2	3	Premature semantic STOP	64.47
LANCE	2	3	PL	-149.50
LANCE	2	3	Queries	1.62
LANCE	3	1	OSR	-36.53
LANCE	3	1	Premature semantic STOP	69.66
LANCE	3	1	PL	-155.92
LANCE	3	1	Queries	2.47
LANCE	3	2	OSR	-28.34
LANCE	3	2	Premature semantic STOP	56.69
LANCE	3	2	PL	-143.70
LANCE	3	2	Queries	2.36
LANCE	3	3	OSR	-18.76
LANCE	3	3	Premature semantic STOP	45.51
LANCE	3	3	PL	-133.49
LANCE	3	3	Queries	2.22
Qwen	1	1	OSR	-42.12
Qwen	1	1	Premature semantic STOP	69.86
Qwen	1	1	PL	-142.60
Qwen	1	1	Queries	4.62
Qwen	1	2	OSR	-38.32
Qwen	1	2	Premature semantic STOP	66.47
Qwen	1	2	PL	-145.03
Qwen	1	2	Queries	2.64
Qwen	1	3	OSR	-39.32

Continued

Judger	K	q	Metric	Difference
Qwen	1	3	Premature semantic STOP	64.47
Qwen	1	3	PL	-139.34
Qwen	1	3	Queries	2.10
Qwen	2	1	OSR	-35.33
Qwen	2	1	Premature semantic STOP	59.28
Qwen	2	1	PL	-127.41
Qwen	2	1	Queries	6.49
Qwen	2	2	OSR	-26.95
Qwen	2	2	Premature semantic STOP	48.90
Qwen	2	2	PL	-119.37
Qwen	2	2	Queries	4.09
Qwen	2	3	OSR	-21.16
Qwen	2	3	Premature semantic STOP	40.12
Qwen	2	3	PL	-111.90
Qwen	2	3	Queries	3.15
Qwen	3	1	OSR	-26.75
Qwen	3	1	Premature semantic STOP	48.50
Qwen	3	1	PL	-117.78
Qwen	3	1	Queries	7.82
Qwen	3	2	OSR	-18.16
Qwen	3	2	Premature semantic STOP	34.73
Qwen	3	2	PL	-101.24
Qwen	3	2	Queries	5.15
Qwen	3	3	OSR	-11.78
Qwen	3	3	Premature semantic STOP	29.14
Qwen	3	3	PL	-87.26
Qwen	3	3	Queries	4.11
SenseNova	1	1	OSR	-41.32
SenseNova	1	1	Premature semantic STOP	71.06
SenseNova	1	1	PL	-144.25
SenseNova	1	1	Queries	3.87
SenseNova	1	2	OSR	-40.32
SenseNova	1	2	Premature semantic STOP	69.46
SenseNova	1	2	PL	-139.80
SenseNova	1	2	Queries	2.67
SenseNova	1	3	OSR	-38.12
SenseNova	1	3	Premature semantic STOP	68.86
SenseNova	1	3	PL	-137.98
SenseNova	1	3	Queries	2.11
SenseNova	2	1	OSR	-32.34
SenseNova	2	1	Premature semantic STOP	57.88
SenseNova	2	1	PL	-125.66
SenseNova	2	1	Queries	6.27
SenseNova	2	2	OSR	-24.95
SenseNova	2	2	Premature semantic STOP	47.90
SenseNova	2	2	PL	-119.22
SenseNova	2	2	Queries	3.87
SenseNova	2	3	OSR	-20.96
SenseNova	2	3	Premature semantic STOP	41.52
SenseNova	2	3	PL	-106.77
SenseNova	2	3	Queries	3.29
SenseNova	3	1	OSR	-25.75
SenseNova	3	1	Premature semantic STOP	47.90
SenseNova	3	1	PL	-116.31
SenseNova	3	1	Queries	7.49
SenseNova	3	2	OSR	-17.17
SenseNova	3	2	Premature semantic STOP	36.53
SenseNova	3	2	PL	-104.28
SenseNova	3	2	Queries	4.64
SenseNova	3	3	OSR	-11.38
SenseNova	3	3	Premature semantic STOP	27.94
SenseNova	3	3	PL	-81.31

Continued

Judger	K	q	Metric	Difference
SenseNova	3	3	Queries	4.40

Table 49: UE486 outside: paired differences from the same judger and q at $K = 1$. ΔSR and ΔSPL are in percentage points. F$\rightarrow$S and S$\rightarrow$F denote failure-to-success and success-to-failure transitions from $K = 1$ to $K = 2, 3$. Counts are listed in repetition order (1, 2, 3).

Judger	K	q	ΔSR	ΔSPL	F$\rightarrow$S (repetitions 1–3)	S$\rightarrow$F (repetitions 1–3)
LANCE	2	1	9.51	9.51	35/37/38	0/0/1
LANCE	2	2	17.02	16.98	65/65/65	0/0/0
LANCE	2	3	24.96	24.29	95/97/95	1/0/0
LANCE	3	1	17.10	17.06	65/66/65	0/0/0
LANCE	3	2	30.02	27.30	114/116/114	0/0/0
LANCE	3	3	25.04	21.98	96/98/94	1/0/0
Qwen	2	1	10.47	10.34	44/45/48	6/5/6
Qwen	2	2	15.53	14.96	62/70/65	7/5/7
Qwen	2	3	15.45	14.32	69/68/72	8/11/13
Qwen	3	1	15.01	14.73	69/63/67	13/5/9
Qwen	3	2	21.03	18.66	95/96/90	13/13/14
Qwen	3	3	12.91	11.07	72/61/72	20/18/19
SenseNova	2	1	11.17	11.06	38/51/47	5/1/2
SenseNova	2	2	15.27	14.59	64/57/60	1/2/3
SenseNova	2	3	17.98	17.25	78/77/84	11/11/11
SenseNova	3	1	16.67	16.18	64/70/65	5/1/2
SenseNova	3	2	20.94	18.70	92/89/92	9/13/11
SenseNova	3	3	15.45	13.53	71/71/79	15/14/15

Table 50: UE486 outside: behavior differences from same-judger/q $K = 1$. Differences in OSR and Premature semantic STOP are in percentage points; executed path length (PL) differences are in meters. Query differences are differences in mean logical queries per episode. Premature semantic STOP is the percentage of episodes terminated by semantic STOP before ever entering the goal region; native STOP is excluded.

Judger	K	q	Metric	Difference
LANCE	2	1	OSR	9.60
LANCE	2	1	Premature semantic STOP	-10.73
LANCE	2	1	PL	6.60
LANCE	2	1	Queries	0.80
LANCE	2	2	OSR	17.02
LANCE	2	2	Premature semantic STOP	-21.03
LANCE	2	2	PL	13.16
LANCE	2	2	Queries	0.77
LANCE	2	3	OSR	25.13
LANCE	2	3	Premature semantic STOP	-29.23
LANCE	2	3	PL	19.92
LANCE	2	3	Queries	0.77
LANCE	3	1	OSR	17.10
LANCE	3	1	Premature semantic STOP	-20.16
LANCE	3	1	PL	13.17
LANCE	3	1	Queries	1.56
LANCE	3	2	OSR	30.10
LANCE	3	2	Premature semantic STOP	-34.55
LANCE	3	2	PL	26.18
LANCE	3	2	Queries	1.51

Continued

Judger	K	q	Metric	Difference
LANCE	3	3	OSR	36.30
LANCE	3	3	Premature semantic STOP	-40.84
LANCE	3	3	PL	38.76
LANCE	3	3	Queries	1.48
Qwen	2	1	OSR	11.52
Qwen	2	1	Premature semantic STOP	-12.48
Qwen	2	1	PL	10.37
Qwen	2	1	Queries	1.27
Qwen	2	2	OSR	17.28
Qwen	2	2	Premature semantic STOP	-19.37
Qwen	2	2	PL	17.13
Qwen	2	2	Queries	0.96
Qwen	2	3	OSR	20.51
Qwen	2	3	Premature semantic STOP	-21.90
Qwen	2	3	PL	21.57
Qwen	2	3	Queries	0.84
Qwen	3	1	OSR	18.41
Qwen	3	1	Premature semantic STOP	-20.86
Qwen	3	1	PL	19.21
Qwen	3	1	Queries	2.30
Qwen	3	2	OSR	28.01
Qwen	3	2	Premature semantic STOP	-30.72
Qwen	3	2	PL	32.93
Qwen	3	2	Queries	1.91
Qwen	3	3	OSR	30.28
Qwen	3	3	Premature semantic STOP	-33.07
Qwen	3	3	PL	48.47
Qwen	3	3	Queries	1.86
SenseNova	2	1	OSR	12.04
SenseNova	2	1	Premature semantic STOP	-13.26
SenseNova	2	1	PL	10.74
SenseNova	2	1	Queries	1.31
SenseNova	2	2	OSR	16.32
SenseNova	2	2	Premature semantic STOP	-17.89
SenseNova	2	2	PL	18.65
SenseNova	2	2	Queries	1.13
SenseNova	2	3	OSR	21.82
SenseNova	2	3	Premature semantic STOP	-24.35
SenseNova	2	3	PL	26.46
SenseNova	2	3	Queries	1.00
SenseNova	3	1	OSR	18.76
SenseNova	3	1	Premature semantic STOP	-21.47
SenseNova	3	1	PL	23.70
SenseNova	3	1	Queries	2.74
SenseNova	3	2	OSR	26.96
SenseNova	3	2	Premature semantic STOP	-29.58
SenseNova	3	2	PL	34.14
SenseNova	3	2	Queries	2.00
SenseNova	3	3	OSR	32.20
SenseNova	3	3	Premature semantic STOP	-35.86
SenseNova	3	3	PL	52.69
SenseNova	3	3	Queries	1.99

Table 51: AirSim180 outside: paired differences from the same judger and q at $K = 1$. ΔSR and ΔSPL are in percentage points. $\mathrm{F} \to \mathrm{S}$ and $\mathrm{S} \to \mathrm{F}$ denote failure-to-success and success-to-failure transitions from $K = 1$ to $K = 2, 3$. Counts are listed in repetition order (1, 2, 3).

Judger	K	q	ΔSR	ΔSPL	$\mathrm{F} \to \mathrm{S}$ (repetitions 1–3)	$\mathrm{S} \to \mathrm{F}$ (repetitions 1–3)
LANCE	2	1	4.79	4.79	8/8/8	0/0/0
LANCE	2	2	9.98	9.98	17/16/17	0/0/0
LANCE	2	3	14.97	14.93	24/26/25	0/0/0
LANCE	3	1	10.78	10.78	18/18/18	0/0/0
LANCE	3	2	18.96	18.31	32/34/29	0/0/0
LANCE	3	3	26.95	25.48	42/44/49	0/0/0
Qwen	2	1	4.99	4.92	12/11/10	2/4/2
Qwen	2	2	8.98	9.04	19/16/20	2/5/3
Qwen	2	3	14.17	13.56	27/28/25	4/3/2
Qwen	3	1	12.77	12.54	25/23/23	2/3/2
Qwen	3	2	16.37	15.53	34/31/29	3/7/2
Qwen	3	3	18.36	16.65	33/36/32	4/2/3
SenseNova	2	1	7.39	7.34	15/11/15	1/2/1
SenseNova	2	2	12.97	12.74	26/21/25	3/2/2
SenseNova	2	3	14.37	14.10	30/24/28	3/3/4
SenseNova	3	1	13.97	13.72	20/28/27	1/3/1
SenseNova	3	2	19.16	18.13	34/34/35	1/4/2
SenseNova	3	3	18.96	17.30	36/35/35	2/3/6

Table 52: AirSim180 outside: behavior differences from same-judger/q $K = 1$. Differences in OSR and Premature semantic STOP are in percentage points; executed path length (PL) differences are in meters. Query differences are differences in mean logical queries per episode. Premature semantic STOP is the percentage of episodes terminated by semantic STOP before ever entering the goal region; native STOP is excluded.

Judger	K	q	Metric	Difference
LANCE	2	1	OSR	4.79
LANCE	2	1	Premature semantic STOP	-7.78
LANCE	2	1	PL	6.83
LANCE	2	1	Queries	0.82
LANCE	2	2	OSR	9.98
LANCE	2	2	Premature semantic STOP	-13.37
LANCE	2	2	PL	13.45
LANCE	2	2	Queries	0.80
LANCE	2	3	OSR	14.97
LANCE	2	3	Premature semantic STOP	-20.56
LANCE	2	3	PL	19.86
LANCE	2	3	Queries	0.77
LANCE	3	1	OSR	10.78
LANCE	3	1	Premature semantic STOP	-16.17
LANCE	3	1	PL	13.44
LANCE	3	1	Queries	1.61
LANCE	3	2	OSR	18.96
LANCE	3	2	Premature semantic STOP	-26.95
LANCE	3	2	PL	25.66
LANCE	3	2	Queries	1.52
LANCE	3	3	OSR	28.54
LANCE	3	3	Premature semantic STOP	-39.52
LANCE	3	3	PL	35.87
LANCE	3	3	Queries	1.37
Qwen	2	1	OSR	6.79
Qwen	2	1	Premature semantic STOP	-10.58
Qwen	2	1	PL	15.19

Continued

Judger	K	q	Metric	Difference
Qwen	2	1	Queries	1.87
Qwen	2	2	OSR	11.38
Qwen	2	2	Premature semantic STOP	-17.56
Qwen	2	2	PL	25.66
Qwen	2	2	Queries	1.45
Qwen	2	3	OSR	18.16
Qwen	2	3	Premature semantic STOP	-24.35
Qwen	2	3	PL	27.43
Qwen	2	3	Queries	1.05
Qwen	3	1	OSR	15.37
Qwen	3	1	Premature semantic STOP	-21.36
Qwen	3	1	PL	24.83
Qwen	3	1	Queries	3.19
Qwen	3	2	OSR	20.16
Qwen	3	2	Premature semantic STOP	-31.74
Qwen	3	2	PL	43.79
Qwen	3	2	Queries	2.50
Qwen	3	3	OSR	27.54
Qwen	3	3	Premature semantic STOP	-35.33
Qwen	3	3	PL	52.07
Qwen	3	3	Queries	2.01
SenseNova	2	1	OSR	8.98
SenseNova	2	1	Premature semantic STOP	-13.17
SenseNova	2	1	PL	18.58
SenseNova	2	1	Queries	2.40
SenseNova	2	2	OSR	15.37
SenseNova	2	2	Premature semantic STOP	-21.56
SenseNova	2	2	PL	20.58
SenseNova	2	2	Queries	1.20
SenseNova	2	3	OSR	17.17
SenseNova	2	3	Premature semantic STOP	-27.35
SenseNova	2	3	PL	31.20
SenseNova	2	3	Queries	1.18
SenseNova	3	1	OSR	15.57
SenseNova	3	1	Premature semantic STOP	-23.15
SenseNova	3	1	PL	27.94
SenseNova	3	1	Queries	3.61
SenseNova	3	2	OSR	23.15
SenseNova	3	2	Premature semantic STOP	-32.93
SenseNova	3	2	PL	35.53
SenseNova	3	2	Queries	1.98
SenseNova	3	3	OSR	26.75
SenseNova	3	3	Premature semantic STOP	-40.92
SenseNova	3	3	PL	56.67
SenseNova	3	3	Queries	2.29

F OUTCOME DECOMPOSITION

The outcome decomposition separates final success, reached but failed, premature semantic STOP, and other non-arrival. Baseline's reached-but-failed share on outside UE486 is 47.29%, compared with 25.75% on outside AirSim180. These are realized failures after region entry; they do not imply that positive visibility judgments identify the correct stop time. Baseline has zero premature semantic STOPs because it does not use the semantic termination module. Native STOP can still end an episode outside the goal region.

Table 53: UE486 outside: mutually exclusive outcome percentages averaged over repetitions. Other non-arrival includes native STOP or limit termination before ever entering the goal region. Premature semantic STOP is the percentage of episodes terminated by semantic STOP before ever entering the goal region; native STOP is excluded.

Judger	K	q	Final success	Reached but failed	Premature semantic STOP	Other non-arrival
Baseline	–	–	11.52	47.29	0.00	41.19
LANCE	1	1	0.09	0.00	78.97	20.94
LANCE	1	2	0.00	0.00	79.84	20.16
LANCE	1	3	0.09	0.00	79.84	20.07
LANCE	2	1	9.60	0.09	68.24	22.08
LANCE	2	2	17.02	0.00	58.81	24.17
LANCE	2	3	25.04	0.17	50.61	24.17
LANCE	3	1	17.19	0.00	58.81	24.00
LANCE	3	2	30.02	0.09	45.29	24.61
LANCE	3	3	25.13	11.26	39.01	24.61
Qwen	1	1	6.02	2.97	66.67	24.35
Qwen	1	2	6.63	3.84	64.66	24.87
Qwen	1	3	8.90	4.36	61.52	25.22
Qwen	2	1	16.49	4.01	54.19	25.31
Qwen	2	2	22.16	5.58	45.29	26.96
Qwen	2	3	24.35	9.42	39.62	26.61
Qwen	3	1	21.03	6.37	45.81	26.79
Qwen	3	2	27.66	10.82	33.94	27.57
Qwen	3	3	21.82	21.73	28.45	28.01
SenseNova	1	1	3.23	2.71	69.98	24.08
SenseNova	1	2	4.80	3.58	67.10	24.52
SenseNova	1	3	5.85	3.93	66.14	24.08
SenseNova	2	1	14.40	3.58	56.72	25.31
SenseNova	2	2	20.07	4.62	49.21	26.09
SenseNova	2	3	23.82	7.77	41.80	26.61
SenseNova	3	1	19.90	4.80	48.52	26.79
SenseNova	3	2	25.74	9.60	37.52	27.14
SenseNova	3	3	21.29	20.68	30.28	27.75

Table 54: AirSim180 outside: mutually exclusive outcome percentages averaged over repetitions. Other non-arrival includes native STOP or limit termination before ever entering the goal region. Premature semantic STOP is the percentage of episodes terminated by semantic STOP before ever entering the goal region; native STOP is excluded.

Judger	K	q	Final success	Reached but failed	Premature semantic STOP	Other non-arrival
Baseline	–	–	21.56	25.75	0.00	52.69
LANCE	1	1	0.00	0.00	85.83	14.17
LANCE	1	2	0.00	0.00	83.63	16.37
LANCE	1	3	0.00	0.00	85.03	14.97
LANCE	2	1	4.79	0.00	78.04	17.17
LANCE	2	2	9.98	0.00	70.26	19.76
LANCE	2	3	14.97	0.00	64.47	20.56
LANCE	3	1	10.78	0.00	69.66	19.56
LANCE	3	2	18.96	0.00	56.69	24.35
LANCE	3	3	26.95	1.60	45.51	25.95
Qwen	1	1	2.79	2.40	69.86	24.95
Qwen	1	2	5.79	3.19	66.47	24.55
Qwen	1	3	5.39	2.59	64.47	27.54
Qwen	2	1	7.78	4.19	59.28	28.74
Qwen	2	2	14.77	5.59	48.90	30.74
Qwen	2	3	19.56	6.59	40.12	33.73
Qwen	3	1	15.57	4.99	48.50	30.94

Continued

Judger	K	q	Final success	Reached but failed	Premature semantic STOP	Other non-arrival
Qwen	3	2	22.16	6.99	34.73	36.13
Qwen	3	3	23.75	11.78	29.14	35.33
SenseNova	1	1	2.20	3.79	71.06	22.95
SenseNova	1	2	3.59	3.39	69.46	23.55
SenseNova	1	3	4.99	4.19	68.86	21.96
SenseNova	2	1	9.58	5.39	57.88	27.15
SenseNova	2	2	16.57	5.79	47.90	29.74
SenseNova	2	3	19.36	6.99	41.52	32.14
SenseNova	3	1	16.17	5.39	47.90	30.54
SenseNova	3	2	22.75	7.39	36.53	33.33
SenseNova	3	3	23.95	11.98	27.94	36.13

G INDIVIDUAL REPETITIONS

SR and SPL are reported for each of the three repetitions for every condition and route group. Their arithmetic means are the values reported above for the complete (K, q) hyperparameter grid. Repetition indices identify paired executions across stopping conditions.

Table 55: UE100, all 100 routes: individual-repetition SR and SPL percentages.

Judger	K	q	SR1	SR2	SR3	SPL1	SPL2	SPL3
Baseline	–	–	22.00	21.00	23.00	20.59	19.13	21.03
LANCE	1	1	30.00	30.00	30.00	30.00	30.00	30.00
LANCE	1	2	30.00	30.00	30.00	30.00	30.00	30.00
LANCE	1	3	30.00	30.00	30.00	30.00	30.00	30.00
LANCE	2	1	32.00	33.00	33.00	25.92	26.89	26.92
LANCE	2	2	41.00	41.00	41.00	30.89	30.24	30.18
LANCE	2	3	37.00	36.00	37.00	32.32	30.33	32.54
LANCE	3	1	42.00	42.00	41.00	30.99	32.02	29.98
LANCE	3	2	34.00	33.00	33.00	31.86	31.41	30.35
LANCE	3	3	33.00	35.00	34.00	28.84	30.85	29.66
Qwen	1	1	31.00	31.00	31.00	29.94	29.12	29.34
Qwen	1	2	32.00	32.00	31.00	30.06	28.62	28.41
Qwen	1	3	35.00	32.00	33.00	33.68	29.72	31.32
Qwen	2	1	35.00	35.00	34.00	26.80	27.43	26.28
Qwen	2	2	42.00	41.00	40.00	32.71	32.82	31.69
Qwen	2	3	36.00	35.00	33.00	33.07	32.53	31.39
Qwen	3	1	40.00	38.00	38.00	30.95	29.16	28.67
Qwen	3	2	37.00	36.00	36.00	35.08	33.18	34.19
Qwen	3	3	32.00	32.00	32.00	27.72	28.28	27.53
SenseNova	1	1	32.00	32.00	31.00	31.28	30.82	30.00
SenseNova	1	2	32.00	35.00	33.00	29.11	34.88	29.34
SenseNova	1	3	30.00	30.00	29.00	30.00	28.86	28.15
SenseNova	2	1	35.00	36.00	34.00	28.06	29.78	27.18
SenseNova	2	2	41.00	37.00	39.00	31.38	28.75	28.74
SenseNova	2	3	37.00	36.00	36.00	32.84	32.48	32.44
SenseNova	3	1	37.00	36.00	38.00	28.06	26.78	29.64
SenseNova	3	2	32.00	31.00	32.00	29.77	29.43	30.35
SenseNova	3	3	31.00	33.00	28.00	27.89	29.51	23.82

Table 56: UE100, 30 inside routes: individual-repetition SR and SPL percentages.

Judger	K	q	SR1	SR2	SR3	SPL1	SPL2	SPL3
Baseline	–	–	40.00	40.00	43.33	36.09	35.48	37.93
LANCE	1	1	100.00	100.00	100.00	100.00	100.00	100.00
LANCE	1	2	100.00	100.00	100.00	100.00	100.00	100.00
LANCE	1	3	100.00	100.00	100.00	100.00	100.00	100.00
LANCE	2	1	100.00	100.00	100.00	79.75	79.65	79.75
LANCE	2	2	100.00	100.00	100.00	66.28	64.12	63.93
LANCE	2	3	63.33	63.33	66.67	47.83	44.92	51.91
LANCE	3	1	100.00	100.00	100.00	63.31	66.73	63.25
LANCE	3	2	43.33	40.00	36.67	39.74	37.38	31.75
LANCE	3	3	40.00	40.00	40.00	36.09	36.09	35.48
Qwen	1	1	96.67	96.67	93.33	93.13	90.39	89.18
Qwen	1	2	90.00	96.67	93.33	84.81	86.32	86.00
Qwen	1	3	93.33	93.33	86.67	89.84	86.66	82.46
Qwen	2	1	93.33	93.33	90.00	66.98	69.38	65.80
Qwen	2	2	83.33	80.00	83.33	53.17	53.14	56.06
Qwen	2	3	43.33	43.33	40.00	37.15	37.46	36.69
Qwen	3	1	83.33	80.00	80.00	54.29	51.46	49.91
Qwen	3	2	43.33	46.67	36.67	41.02	41.27	34.52
Qwen	3	3	43.33	40.00	36.67	37.32	36.09	31.22
SenseNova	1	1	93.33	96.67	90.00	91.33	93.17	87.78
SenseNova	1	2	93.33	96.67	96.67	84.09	96.67	84.86
SenseNova	1	3	93.33	90.00	90.00	93.33	86.20	87.17
SenseNova	2	1	96.67	96.67	96.67	73.53	76.31	74.89
SenseNova	2	2	90.00	83.33	93.33	58.04	56.32	60.37
SenseNova	2	3	56.67	56.67	56.67	43.78	46.71	46.76
SenseNova	3	1	83.33	86.67	86.67	53.53	56.67	59.78
SenseNova	3	2	40.00	40.00	36.67	36.97	37.86	34.92
SenseNova	3	3	36.67	40.00	40.00	34.52	36.09	34.60

Table 57: UE100, 70 outside routes: individual-repetition SR and SPL percentages.

Judger	K	q	SR1	SR2	SR3	SPL1	SPL2	SPL3
Baseline	–	–	14.29	12.86	14.29	13.94	12.12	13.78
LANCE	1	1	0.00	0.00	0.00	0.00	0.00	0.00
LANCE	1	2	0.00	0.00	0.00	0.00	0.00	0.00
LANCE	1	3	0.00	0.00	0.00	0.00	0.00	0.00
LANCE	2	1	2.86	4.29	4.29	2.86	4.29	4.29
LANCE	2	2	15.71	15.71	15.71	15.71	15.71	15.71
LANCE	2	3	25.71	24.29	24.29	25.67	24.08	24.24
LANCE	3	1	17.14	17.14	15.71	17.14	17.14	15.71
LANCE	3	2	30.00	30.00	31.43	28.48	28.85	29.75
LANCE	3	3	30.00	32.86	31.43	25.73	28.60	27.17
Qwen	1	1	2.86	2.86	4.29	2.86	2.86	3.69
Qwen	1	2	7.14	4.29	4.29	6.59	3.89	3.73
Qwen	1	3	10.00	5.71	10.00	9.60	5.32	9.41
Qwen	2	1	10.00	10.00	10.00	9.58	9.45	9.34
Qwen	2	2	24.29	24.29	21.43	23.94	24.11	21.25
Qwen	2	3	32.86	31.43	30.00	31.32	30.41	29.11
Qwen	3	1	21.43	20.00	20.00	20.95	19.60	19.57
Qwen	3	2	34.29	31.43	35.71	32.54	29.71	34.04
Qwen	3	3	27.14	28.57	30.00	23.60	24.94	25.96
SenseNova	1	1	5.71	4.29	5.71	5.55	4.10	5.24
SenseNova	1	2	5.71	8.57	5.71	5.55	8.41	5.55
SenseNova	1	3	2.86	4.29	2.86	2.86	4.29	2.86
SenseNova	2	1	8.57	10.00	7.14	8.57	9.83	6.73
SenseNova	2	2	20.00	17.14	15.71	19.95	16.94	15.19

Continued

Judger	K	q	SR1	SR2	SR3	SPL1	SPL2	SPL3
SenseNova	2	3	28.57	27.14	27.14	28.15	26.38	26.31
SenseNova	3	1	17.14	14.29	17.14	17.14	13.97	16.72
SenseNova	3	2	28.57	27.14	30.00	26.68	25.82	28.39
SenseNova	3	3	28.57	30.00	22.86	25.05	26.68	19.20

Table 58: UE486, all 486 routes: individual-repetition SR and SPL percentages.

Judger	K	q	SR1	SR2	SR3	SPL1	SPL2	SPL3
Baseline	–	–	17.49	17.08	18.52	15.65	15.07	16.21
LANCE	1	1	21.40	21.40	21.60	21.40	21.40	21.60
LANCE	1	2	21.40	21.40	21.40	21.40	21.40	21.40
LANCE	1	3	21.60	21.40	21.40	21.55	21.40	21.40
LANCE	2	1	28.60	29.01	29.22	24.38	24.72	25.27
LANCE	2	2	34.77	34.77	34.77	27.94	27.36	27.11
LANCE	2	3	34.57	34.77	35.19	29.79	29.92	31.31
LANCE	3	1	34.77	34.98	34.98	28.15	27.79	27.76
LANCE	3	2	33.33	33.95	33.13	29.73	29.94	29.52
LANCE	3	3	28.19	29.63	29.01	24.78	25.83	24.99
Qwen	1	1	24.69	24.49	24.90	24.05	23.43	23.73
Qwen	1	2	25.93	24.69	24.69	24.26	23.61	23.64
Qwen	1	3	26.54	25.93	25.10	25.44	25.42	24.13
Qwen	2	1	32.30	31.48	32.92	28.11	27.27	28.79
Qwen	2	2	35.80	35.60	35.19	29.48	29.74	29.87
Qwen	2	3	33.95	32.92	33.13	29.16	28.57	28.40
Qwen	3	1	34.57	34.77	35.60	29.12	28.94	29.99
Qwen	3	2	32.51	31.48	31.28	28.91	27.37	27.30
Qwen	3	3	27.57	24.90	26.75	24.65	22.03	23.66
SenseNova	1	1	24.49	22.02	22.84	24.05	21.67	21.88
SenseNova	1	2	23.66	24.28	24.69	22.59	22.76	23.76
SenseNova	1	3	24.07	24.28	24.07	23.49	23.33	23.25
SenseNova	2	1	31.48	31.89	31.48	26.99	27.41	27.06
SenseNova	2	2	34.98	33.95	34.77	28.78	27.67	28.72
SenseNova	2	3	31.89	32.51	33.74	27.79	28.15	29.58
SenseNova	3	1	35.19	34.77	34.36	29.49	28.91	28.15
SenseNova	3	2	30.66	29.84	31.28	26.80	26.47	27.27
SenseNova	3	3	24.69	25.93	26.34	21.78	22.76	23.18

Table 59: UE486, 104 inside routes: individual-repetition SR and SPL percentages.

Judger	K	q	SR1	SR2	SR3	SPL1	SPL2	SPL3
Baseline	–	–	38.46	40.38	42.31	32.58	33.08	34.42
LANCE	1	1	100.00	100.00	100.00	100.00	100.00	100.00
LANCE	1	2	100.00	100.00	100.00	100.00	100.00	100.00
LANCE	1	3	100.00	100.00	100.00	100.00	100.00	100.00
LANCE	2	1	100.00	100.00	100.00	80.26	79.96	81.57
LANCE	2	2	100.00	100.00	100.00	68.47	65.34	64.19
LANCE	2	3	70.19	69.23	73.08	50.42	48.99	57.53
LANCE	3	1	100.00	100.00	100.00	69.07	66.39	66.68
LANCE	3	2	46.15	47.12	45.19	38.90	38.86	38.20
LANCE	3	3	39.42	44.23	45.19	34.77	37.79	37.81
Qwen	1	1	92.31	93.27	94.23	90.04	89.32	89.62
Qwen	1	2	94.23	94.23	90.38	87.93	90.08	86.64
Qwen	1	3	91.35	86.54	86.54	87.86	84.67	83.09
Qwen	2	1	91.35	87.50	91.35	72.53	69.64	73.50
Qwen	2	2	87.50	82.69	83.65	61.70	58.80	61.41

Continued

Judger	K	q	SR1	SR2	SR3	SPL1	SPL2	SPL3
Qwen	2	3	67.31	64.42	67.31	50.18	49.06	50.71
Qwen	3	1	84.62	85.58	88.46	60.49	60.69	64.14
Qwen	3	2	46.15	46.15	48.08	39.38	36.58	39.47
Qwen	3	3	46.15	40.38	43.27	39.86	34.60	37.43
SenseNova	1	1	95.19	96.15	97.12	93.27	94.55	92.98
SenseNova	1	2	95.19	95.19	96.15	90.32	88.38	92.70
SenseNova	1	3	91.35	90.38	92.31	89.73	86.97	89.71
SenseNova	2	1	96.15	94.23	94.23	75.45	74.26	73.97
SenseNova	2	2	87.50	87.50	88.46	61.88	61.16	62.61
SenseNova	2	3	63.46	65.38	67.31	48.49	49.04	51.03
SenseNova	3	1	88.46	89.42	90.38	64.22	64.12	62.77
SenseNova	3	2	48.08	48.08	49.04	39.24	39.91	39.57
SenseNova	3	3	40.38	43.27	41.35	34.39	37.05	34.98

Table 60: UE486, 382 outside routes: individual-repetition SR and SPL percentages.

Judger	K	q	SR1	SR2	SR3	SPL1	SPL2	SPL3
Baseline	–	–	11.78	10.73	12.04	11.04	10.16	11.25
LANCE	1	1	0.00	0.00	0.26	0.00	0.00	0.26
LANCE	1	2	0.00	0.00	0.00	0.00	0.00	0.00
LANCE	1	3	0.26	0.00	0.00	0.20	0.00	0.00
LANCE	2	1	9.16	9.69	9.95	9.16	9.69	9.95
LANCE	2	2	17.02	17.02	17.02	16.90	17.02	17.02
LANCE	2	3	24.87	25.39	24.87	24.18	24.72	24.17
LANCE	3	1	17.02	17.28	17.28	17.02	17.28	17.16
LANCE	3	2	29.84	30.37	29.84	27.23	27.52	27.16
LANCE	3	3	25.13	25.65	24.61	22.06	22.58	21.50
Qwen	1	1	6.28	5.76	6.02	6.09	5.49	5.79
Qwen	1	2	7.33	5.76	6.81	6.93	5.51	6.49
Qwen	1	3	8.90	9.42	8.38	8.45	9.29	8.08
Qwen	2	1	16.23	16.23	17.02	16.02	15.74	16.62
Qwen	2	2	21.73	22.77	21.99	20.71	21.82	21.28
Qwen	2	3	24.87	24.35	23.82	23.44	23.00	22.33
Qwen	3	1	20.94	20.94	21.20	20.58	20.29	20.70
Qwen	3	2	28.80	27.49	26.70	26.06	24.86	23.98
Qwen	3	3	22.51	20.68	22.25	20.51	18.61	19.91
SenseNova	1	1	5.24	1.83	2.62	5.21	1.82	2.52
SenseNova	1	2	4.19	4.97	5.24	4.15	4.89	4.99
SenseNova	1	3	5.76	6.28	5.50	5.45	6.00	5.15
SenseNova	2	1	13.87	14.92	14.40	13.80	14.65	14.28
SenseNova	2	2	20.68	19.37	20.16	19.77	18.55	19.49
SenseNova	2	3	23.30	23.56	24.61	22.16	22.46	23.74
SenseNova	3	1	20.68	19.90	19.11	20.04	19.32	18.73
SenseNova	3	2	25.92	24.87	26.44	23.41	22.82	23.93
SenseNova	3	3	20.42	21.20	22.25	18.35	18.87	19.96

Table 61: UE586, all 586 routes: individual-repetition SR and SPL percentages.

Judger	K	q	SR1	SR2	SR3	SPL1	SPL2	SPL3
Baseline	–	–	18.26	17.75	19.28	16.49	15.76	17.03
LANCE	1	1	22.87	22.87	23.04	22.87	22.87	23.04
LANCE	1	2	22.87	22.87	22.87	22.87	22.87	22.87
LANCE	1	3	23.04	22.87	22.87	23.00	22.87	22.87
LANCE	2	1	29.18	29.69	29.86	24.64	25.09	25.56
LANCE	2	2	35.84	35.84	35.84	28.44	27.85	27.63

Continued

Judger	K	q	SR1	SR2	SR3	SPL1	SPL2	SPL3
LANCE	2	3	34.98	34.98	35.49	30.22	29.99	31.52
LANCE	3	1	36.01	36.18	36.01	28.64	28.51	28.14
LANCE	3	2	33.45	33.79	33.11	30.09	30.19	29.66
LANCE	3	3	29.01	30.55	29.86	25.47	26.69	25.79
Qwen	1	1	25.77	25.60	25.94	25.06	24.40	24.68
Qwen	1	2	26.96	25.94	25.77	25.25	24.46	24.46
Qwen	1	3	27.99	26.96	26.45	26.85	26.15	25.36
Qwen	2	1	32.76	32.08	33.11	27.89	27.30	28.36
Qwen	2	2	36.86	36.52	36.01	30.03	30.26	30.18
Qwen	2	3	34.30	33.28	33.11	29.83	29.25	28.91
Qwen	3	1	35.49	35.32	36.01	29.43	28.98	29.77
Qwen	3	2	33.28	32.25	32.08	29.96	28.36	28.47
Qwen	3	3	28.33	26.11	27.65	25.17	23.10	24.32
SenseNova	1	1	25.77	23.72	24.23	25.29	23.23	23.26
SenseNova	1	2	25.09	26.11	26.11	23.70	24.83	24.71
SenseNova	1	3	25.09	25.26	24.91	24.60	24.27	24.08
SenseNova	2	1	32.08	32.59	31.91	27.17	27.81	27.08
SenseNova	2	2	36.01	34.47	35.49	29.22	27.85	28.72
SenseNova	2	3	32.76	33.11	34.13	28.65	28.89	30.07
SenseNova	3	1	35.49	34.98	34.98	29.25	28.55	28.41
SenseNova	3	2	30.89	30.03	31.40	27.30	26.98	27.80
SenseNova	3	3	25.77	27.13	26.62	22.82	23.91	23.29

Table 62: UE586, 134 inside routes: individual-repetition SR and SPL percentages.

Judger	K	q	SR1	SR2	SR3	SPL1	SPL2	SPL3
Baseline	–	–	38.81	40.30	42.54	33.36	33.62	35.21
LANCE	1	1	100.00	100.00	100.00	100.00	100.00	100.00
LANCE	1	2	100.00	100.00	100.00	100.00	100.00	100.00
LANCE	1	3	100.00	100.00	100.00	100.00	100.00	100.00
LANCE	2	1	100.00	100.00	100.00	80.15	79.89	81.16
LANCE	2	2	100.00	100.00	100.00	67.98	65.07	64.14
LANCE	2	3	68.66	67.91	71.64	49.84	48.08	56.27
LANCE	3	1	100.00	100.00	100.00	67.78	66.47	65.91
LANCE	3	2	45.52	45.52	43.28	39.09	38.53	36.76
LANCE	3	3	39.55	43.28	44.03	35.07	37.41	37.29
Qwen	1	1	93.28	94.03	94.03	90.73	89.56	89.52
Qwen	1	2	93.28	94.78	91.04	87.23	89.24	86.50
Qwen	1	3	91.79	88.06	86.57	88.31	85.12	82.95
Qwen	2	1	91.79	88.81	91.04	71.29	69.58	71.77
Qwen	2	2	86.57	82.09	83.58	59.79	57.53	60.21
Qwen	2	3	61.94	59.70	61.19	47.27	46.47	47.58
Qwen	3	1	84.33	84.33	86.57	59.10	58.63	60.96
Qwen	3	2	45.52	46.27	45.52	39.75	37.63	38.36
Qwen	3	3	45.52	40.30	41.79	39.29	34.94	36.04
SenseNova	1	1	94.78	96.27	95.52	92.83	94.24	91.82
SenseNova	1	2	94.78	95.52	96.27	88.93	90.23	90.95
SenseNova	1	3	91.79	90.30	91.79	90.54	86.80	89.14
SenseNova	2	1	96.27	94.78	94.78	75.02	74.72	74.17
SenseNova	2	2	88.06	86.57	89.55	61.02	60.07	62.11
SenseNova	2	3	61.94	63.43	64.93	47.44	48.52	50.07
SenseNova	3	1	87.31	88.81	89.55	61.83	62.45	62.10
SenseNova	3	2	46.27	46.27	46.27	38.73	39.45	38.53
SenseNova	3	3	39.55	42.54	41.04	34.42	36.84	34.89

Table 63: UE586, 452 outside routes: individual-repetition SR and SPL percentages.

Judger	K	q	SR1	SR2	SR3	SPL1	SPL2	SPL3
Baseline	–	–	12.17	11.06	12.39	11.49	10.47	11.64
LANCE	1	1	0.00	0.00	0.22	0.00	0.00	0.22
LANCE	1	2	0.00	0.00	0.00	0.00	0.00	0.00
LANCE	1	3	0.22	0.00	0.00	0.17	0.00	0.00
LANCE	2	1	8.19	8.85	9.07	8.19	8.85	9.07
LANCE	2	2	16.81	16.81	16.81	16.72	16.81	16.81
LANCE	2	3	25.00	25.22	24.78	24.41	24.62	24.18
LANCE	3	1	17.04	17.26	17.04	17.04	17.26	16.94
LANCE	3	2	29.87	30.31	30.09	27.42	27.72	27.56
LANCE	3	3	25.88	26.77	25.66	22.63	23.51	22.38
Qwen	1	1	5.75	5.31	5.75	5.59	5.08	5.46
Qwen	1	2	7.30	5.53	6.42	6.88	5.26	6.07
Qwen	1	3	9.07	8.85	8.63	8.62	8.67	8.29
Qwen	2	1	15.27	15.27	15.93	15.02	14.76	15.49
Qwen	2	2	22.12	23.01	21.90	21.21	22.18	21.27
Qwen	2	3	26.11	25.44	24.78	24.66	24.14	23.38
Qwen	3	1	21.02	20.80	21.02	20.64	20.19	20.52
Qwen	3	2	29.65	28.10	28.10	27.06	25.61	25.54
Qwen	3	3	23.23	21.90	23.45	20.99	19.59	20.85
SenseNova	1	1	5.31	2.21	3.10	5.26	2.18	2.94
SenseNova	1	2	4.42	5.53	5.31	4.37	5.44	5.08
SenseNova	1	3	5.31	5.97	5.09	5.05	5.73	4.80
SenseNova	2	1	13.05	14.16	13.27	12.99	13.91	13.11
SenseNova	2	2	20.58	19.03	19.47	19.80	18.30	18.83
SenseNova	2	3	24.12	24.12	25.00	23.08	23.07	24.14
SenseNova	3	1	20.13	19.03	18.81	19.59	18.49	18.42
SenseNova	3	2	26.33	25.22	26.99	23.91	23.28	24.62
SenseNova	3	3	21.68	22.57	22.35	19.38	20.08	19.85

Table 64: AirSim180, all 180 routes: individual-repetition SR and SPL percentages.

Judger	K	q	SR1	SR2	SR3	SPL1	SPL2	SPL3
Baseline	–	–	20.56	24.44	26.67	19.12	22.87	25.47
LANCE	1	1	7.22	7.22	7.22	7.22	7.22	7.22
LANCE	1	2	7.22	7.22	7.22	7.22	7.22	7.22
LANCE	1	3	7.22	7.22	7.22	7.22	7.22	7.22
LANCE	2	1	11.67	11.67	11.67	11.28	11.28	11.21
LANCE	2	2	16.67	16.11	16.67	15.43	14.41	14.97
LANCE	2	3	20.56	21.11	21.11	18.57	19.27	18.96
LANCE	3	1	17.22	17.22	17.22	15.95	15.99	15.62
LANCE	3	2	22.78	23.89	21.11	21.33	22.78	19.56
LANCE	3	3	27.78	28.33	31.11	26.55	26.79	29.46
Qwen	1	1	8.89	10.56	10.00	8.22	10.34	9.47
Qwen	1	2	11.11	13.33	12.78	11.02	13.00	12.58
Qwen	1	3	12.78	10.56	12.78	12.64	10.40	12.10
Qwen	2	1	14.44	14.44	13.89	13.31	14.06	12.72
Qwen	2	2	20.56	19.44	22.22	18.93	17.53	20.42
Qwen	2	3	25.00	24.44	25.56	22.43	21.90	22.43
Qwen	3	1	21.67	21.67	21.67	19.82	19.58	19.47
Qwen	3	2	25.56	24.44	25.56	24.42	22.81	23.11
Qwen	3	3	26.11	27.22	26.11	23.76	24.87	23.99
SenseNova	1	1	8.33	9.44	8.89	8.16	9.28	8.81
SenseNova	1	2	10.00	11.67	10.00	9.88	11.13	9.84
SenseNova	1	3	11.67	11.67	11.67	11.36	11.54	11.25
SenseNova	2	1	16.11	14.44	17.22	15.61	13.33	16.24
SenseNova	2	2	22.78	22.22	22.78	20.91	20.38	21.33

Continued

Judger	K	q	SR1	SR2	SR3	SPL1	SPL2	SPL3
SenseNova	2	3	26.67	22.78	24.44	24.08	20.24	22.60
SenseNova	3	1	18.89	23.33	23.89	17.69	21.28	21.74
SenseNova	3	2	25.56	26.11	26.11	23.69	24.43	24.12
SenseNova	3	3	27.78	26.11	26.11	26.21	23.48	23.65

Table 65: AirSim180, 13 inside routes: individual-repetition SR and SPL percentages.

Judger	K	q	SR1	SR2	SR3	SPL1	SPL2	SPL3
Baseline	–	–	46.15	53.85	61.54	39.51	46.95	60.63
LANCE	1	1	100.00	100.00	100.00	100.00	100.00	100.00
LANCE	1	2	100.00	100.00	100.00	100.00	100.00	100.00
LANCE	1	3	100.00	100.00	100.00	100.00	100.00	100.00
LANCE	2	1	100.00	100.00	100.00	94.64	94.64	93.73
LANCE	2	2	100.00	100.00	100.00	82.90	76.49	76.49
LANCE	2	3	100.00	92.31	100.00	72.98	67.23	70.90
LANCE	3	1	100.00	100.00	100.00	82.42	82.90	77.78
LANCE	3	2	69.23	69.23	69.23	60.51	60.77	54.58
LANCE	3	3	61.54	53.85	53.85	61.54	53.36	49.51
Qwen	1	1	100.00	100.00	100.00	91.03	100.00	93.59
Qwen	1	2	92.31	100.00	100.00	91.40	100.00	100.00
Qwen	1	3	100.00	92.31	100.00	100.00	92.31	93.59
Qwen	2	1	100.00	100.00	92.31	88.23	94.64	78.94
Qwen	2	2	92.31	100.00	100.00	71.36	76.49	76.01
Qwen	2	3	92.31	92.31	100.00	68.77	65.77	66.56
Qwen	3	1	100.00	100.00	100.00	76.49	75.25	76.49
Qwen	3	2	53.85	69.23	69.23	50.00	60.99	49.20
Qwen	3	3	61.54	61.54	61.54	51.10	57.21	55.13
SenseNova	1	1	92.31	100.00	92.31	92.31	100.00	92.31
SenseNova	1	2	100.00	100.00	100.00	100.00	93.59	100.00
SenseNova	1	3	100.00	100.00	92.31	100.00	100.00	91.40
SenseNova	2	1	92.31	100.00	100.00	86.95	88.23	88.87
SenseNova	2	2	100.00	100.00	100.00	76.87	82.42	82.90
SenseNova	2	3	100.00	92.31	84.62	68.33	64.39	69.06
SenseNova	3	1	92.31	100.00	100.00	77.72	78.16	77.13
SenseNova	3	2	61.54	69.23	69.23	53.30	55.87	58.97
SenseNova	3	3	61.54	53.85	69.23	57.21	47.44	62.64

Table 66: AirSim180, 167 outside routes: individual-repetition SR and SPL percentages.

Judger	K	q	SR1	SR2	SR3	SPL1	SPL2	SPL3
Baseline	–	–	18.56	22.16	23.95	17.54	20.99	22.73
LANCE	1	1	0.00	0.00	0.00	0.00	0.00	0.00
LANCE	1	2	0.00	0.00	0.00	0.00	0.00	0.00
LANCE	1	3	0.00	0.00	0.00	0.00	0.00	0.00
LANCE	2	1	4.79	4.79	4.79	4.79	4.79	4.79
LANCE	2	2	10.18	9.58	10.18	10.18	9.58	10.18
LANCE	2	3	14.37	15.57	14.97	14.34	15.54	14.92
LANCE	3	1	10.78	10.78	10.78	10.78	10.78	10.78
LANCE	3	2	19.16	20.36	17.37	18.28	19.83	16.83
LANCE	3	3	25.15	26.35	29.34	23.82	24.72	27.90
Qwen	1	1	1.80	3.59	2.99	1.78	3.36	2.92
Qwen	1	2	4.79	6.59	5.99	4.76	6.22	5.78
Qwen	1	3	5.99	4.19	5.99	5.84	4.02	5.76
Qwen	2	1	7.78	7.78	7.78	7.47	7.78	7.57

Continued

Judger	K	q	SR1	SR2	SR3	SPL1	SPL2	SPL3
Qwen	2	2	14.97	13.17	16.17	14.85	12.94	16.09
Qwen	2	3	19.76	19.16	19.76	18.82	18.49	18.99
Qwen	3	1	15.57	15.57	15.57	15.41	15.24	15.03
Qwen	3	2	23.35	20.96	22.16	22.43	19.84	21.08
Qwen	3	3	23.35	24.55	23.35	21.63	22.35	21.57
SenseNova	1	1	1.80	2.40	2.40	1.61	2.22	2.31
SenseNova	1	2	2.99	4.79	2.99	2.87	4.71	2.83
SenseNova	1	3	4.79	4.79	5.39	4.46	4.65	5.01
SenseNova	2	1	10.18	7.78	10.78	10.06	7.50	10.59
SenseNova	2	2	16.77	16.17	16.77	16.55	15.55	16.54
SenseNova	2	3	20.96	17.37	19.76	20.63	16.80	18.98
SenseNova	3	1	13.17	17.37	17.96	13.02	16.85	17.42
SenseNova	3	2	22.75	22.75	22.75	21.39	21.99	21.41
SenseNova	3	3	25.15	23.95	22.75	23.80	21.62	20.62