BREAKING SUCCESS-PATH COLLAPSE: ADVANCING AERIAL VISION-AND-LANGUAGE NAVIGATION VIA ACTIVE ENVIRONMENT SHAPING

Anonymous authors
Paper under double-blind review

ABSTRACT

Simulation-based trajectory synthesis has emerged as a promising route towards scalable training data for aerial vision-and-language navigation. However, existing generators, whether search-based planning or reinforcement learning algorithms, share a failure mode that we identify and term *Success-Path Collapse*. Specifically, trajectories converge onto a narrow set of near-optimal routes. This behavioral redundancy, masked by high task success, silently propagates into downstream training data and compromises the robustness of learned navigation policies. In this work, we propose a novel perspective to actively restructure the environment to elicit multimodal trajectory generation, while remaining fully compatible with existing data synthesis pipelines. We instantiate this perspective with Trajectory-Aware Environment Perturbation (TAEP), which diagnoses collapse structures from successful trajectories, including biased junction exits, over-visited hotspots, and redundant route clusters, and places targeted virtual obstacles that naturally turn the pursuit of successful trajectories into continual route discovery. Experiments in AirSim urban scenes show that TAEP achieves the strongest overall trajectory diversity across a suite of diagnostic metrics, outperforming vanilla PPO, intrinsic-reward exploration, and random perturbation. TAEP-generated demonstrations translate these gains into improved closed-loop VLA navigation, more than doubling vanilla PPO's success rate on the longest tasks and improving mean success rate by 6.5 percentage points over the strongest evaluated baseline. Code and data are available at `https://anonymous.4open.science/r/TAEP-ICLR-22E1/`.

1 INTRODUCTION

Aerial vision-and-language navigation (VLN; Liu et al., 2023) requires an unmanned aerial vehicle (UAV) to reach a language-specified target in continuous airspace, supporting applications such as disaster rescue, infrastructure inspection, and security patrol. Training reliable aerial vision-language-action (VLA) models (Zitkovich et al., 2023; Kim et al., 2024) faces two major data challenges. **(1) High collection costs.** Manual piloting makes flight-data collection costly, while airspace regulations further constrain its scalability in real-world settings (Liu et al., 2023; Wang et al., 2025a). **(2) Long-horizon distribution shift.** When a VLA model predicts actions autoregressively over hundreds of meters, small errors accumulate and carry the policy beyond the state distribution represented in its demonstrations (Ross et al., 2011). Effective training data must therefore capture a broad range of spatial contexts, turning patterns, and route structures to support reliable navigation under these deviations.

Simulation-based synthesis offers a scalable source of flight demonstrations (Gao et al., 2025; Li et al., 2026). A generator with privileged access to scene geometry first constructs collision-free trajectories between sampled start-goal pairs. Visual observations are then rendered along these trajectories and paired with subsequently synthesized instructions to fine-tune a VLA model. This automation removes much of the collection burden but leaves the challenge of behavioral coverage to the trajectory generator. Existing pipelines use search-based planners (Hart et al., 1968), including action-primitive search in OpenFly (Gao et al., 2025) and dynamically feasible planning in FlyMirage (Li et al., 2026). Their preference for efficient paths yields near-deterministic routes under fixed

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.

planning criteria. Reinforcement learning (RL) offers an alternative to planning by learning navigation end-to-end from reward. However, maximizing task return also favors convergence to a narrow set of high-return routes. A PPO generator (Schulman et al., 2017) that we train in an AirSim urban scene achieves an 89% success rate, yet repeated rollouts for the same start-goal pair follow nearly identical routes. We identify this phenomenon as **Success-Path Collapse**, a bottleneck in which successful trajectory generation produces large but behaviorally narrow corpora. This redundancy propagates into the demonstrations used for downstream VLA training.

Intrinsic rewards encourage exploration and diverse behavior through visitation counts (Strehl & Littman, 2008), prediction error (Pathak et al., 2017; Burda et al., 2018), or skill discrimination (Eysenbach et al., 2018). For navigation data generation, their effectiveness hinges on two coupled design problems. The diversity signal must distinguish meaningful route alternatives from variations that contribute little new behavior. Novelty bonuses can reward local deviations while leaving the dominant route unchanged. The auxiliary objective must also exert enough influence to overcome route preferences without sacrificing successful, efficient navigation. Downstream training data quality is therefore sensitive to both reward design and weight calibration. Here, as an alternative, we shift the intervention to environment topology, so that the pursuit of navigation success is naturally translated into the discovery of diverse routes.

We instantiate this approach with **Trajectory-Aware Environment Perturbation (TAEP)**, a framework that converts evidence of trajectory collapse into targeted changes in environment topology. TAEP diagnoses biased junction choices, visitation hotspots, and repeated route segments in successful trajectories. These diagnostics guide virtual obstacle placement to reshape local traversability and redirect the policy toward underexplored routes. Fine-tuning in the modified environment turns the pursuit of task success into continued route discovery under the existing navigation objective. Replaying the resulting trajectories produces demonstrations spanning a richer range of locations, viewpoints, and maneuvers for aerial VLN. Experiments in two AirSim urban scenes show that TAEP achieves the strongest overall trajectory diversity across five diagnostic metrics, outperforming vanilla PPO, an intrinsic-reward baseline (MBIE-EB), and geometry-matched random perturbation. With training corpora matched in scale, these gains transfer to closed-loop VLA navigation, improving success rate, path efficiency, and navigation error over all three baselines. TAEP demonstrations more than double vanilla PPO's success rate on the longest tasks, from 34.4% to 71.1%, and raise mean success rate across distance groups by 6.5 percentage points over the strongest evaluated baseline.

Our main contributions are summarized below.

1. We identify *Success-Path Collapse* in aerial VLN trajectory generation and introduce a unified diagnostic and evaluation suite spanning decision, traffic, and route-level behavior.
2. We propose TAEP, a unified pipeline linking trajectory diagnosis, targeted environment shaping, and policy adaptation to diverse demonstrations for VLA training.
3. Experiments in AirSim urban scenes demonstrate consistent gains in trajectory diversity and downstream VLA navigation, raising mean success rate by 6.5 percentage points over the strongest baseline and more than doubling vanilla PPO's long-range success rate.

2 Related Work

Aerial VLN and automated data generation. VLN requires an embodied agent to navigate to a destination by grounding language instructions in visual observations. It has expanded from indoor (Anderson et al., 2018; Krantz et al., 2020) and outdoor ground navigation (Chen et al., 2019) to UAV navigation in open airspace (Liu et al., 2023; Wang et al., 2025b). VLA models support this expansion through the visual-semantic generalization acquired from large-scale pretraining (Zitkovich et al., 2023; Kim et al., 2024; Gao et al., 2025), but adapting them to aerial control requires extensive flight demonstrations. To meet this demand, OpenFly (Gao et al., 2025) and FlyMirage (Li et al., 2026) automate trajectory synthesis and visual-language data construction. However, scaling these pipelines also propagates the generator's route-selection bias into the demonstrations and, through fine-tuning, into the VLA policy. TAEP addresses this bias by diversifying trajectory generation.

Data augmentation for navigation. Imitation policies accumulate errors when their actions lead beyond the state distribution covered by demonstrations (Ross et al., 2011). To reduce this mis-

match, data augmentation broadens training coverage through instruction synthesis (Fried et al., 2018; Tan et al., 2019), visual environment editing (Li et al., 2022), recombination of scenes and paths (Liu et al., 2021), and large-scale scene-based generation (Wang et al., 2023; Zhong et al., 2024). However, these approaches primarily enrich language, appearance, and scene composition. Within a scene, repeatedly selecting the same successful routes restricts the navigation situations demonstrated even when their descriptions and renderings vary. TAEP therefore targets the trajectory distribution itself, expanding the locations, viewpoints, and maneuvers represented in the corpus through environment shaping.

Exploration and behavioral diversity in RL. Diversity can also be encouraged during policy learning through visitation bonuses (Strehl & Littman, 2008), prediction-error rewards (Pathak et al., 2017; Burda et al., 2018), and skill discovery (Eysenbach et al., 2018). Applying these objectives to demonstration generation, however, requires aligning the rewarded variation with useful successful routes. Local deviations can increase novelty while leaving dominant paths unchanged, so the diversity signal must distinguish these deviations from meaningful route alternatives. Its weight also requires careful tuning to balance diversification with successful, efficient navigation, making downstream data quality sensitive to both reward design and calibration. TAEP offers a more natural way to induce route diversity by reshaping the environment at diagnosed concentrations of successful behavior, so that continued navigation learning itself drives the discovery of new routes.

Environment design and curricula. Environment design shapes learning by controlling the tasks and conditions under which policies are trained. POET co-evolves environments and their solutions (Wang et al., 2019), while PAIRED (Dennis et al., 2020) and ACCEL (Parker-Holder et al., 2022) use regret to construct adaptive curricula. These works primarily use environment variation to develop policy capabilities across tasks, rather than to broaden behavioral coverage. A policy can solve a task reliably while repeatedly following the same route, leaving substantial redundancy in the data it generates. TAEP therefore selects environment edits from this redundancy, directing policy adaptation toward a richer distribution of successful demonstrations for downstream learning.

3 Diagnosing Success-Path Collapse

We examine successful trajectories from a PPO generator trained on an urban occupancy map derived from AirSim point clouds (Section 5.1). The policy reaches an 89% success rate, yet its trajectories repeatedly visit the same parts of the scene and follow similar routes. We refer to this concentration of successful behavior as **Success-Path Collapse**.

3.1 Three Levels of Collapse

Figure 1 illustrates three levels of collapse.

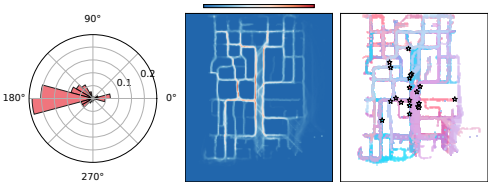
Decision level: branching bias. At junctions with multiple traversable exits, the policy repeatedly favors particular directions.

Traffic level: spatial concentration. Visits concentrate on a small part of the road network, with limited use of other traversable regions.

Route level: backbone redundancy. Many trajectories share long segments, forming a small set of recurring route patterns.

These aspects connect local choices to corpus-wide structure: repeated exit choices can concentrate traffic, and frequently used segments can recur across complete routes.

They provide complementary locations and spatial scales for intervention.



(a) Directional bias (b) Visitation concentration (c) Route redundancy

Figure 1: Three views of Success-Path Collapse. (a) Normalized heading frequency reveals directional bias. (b) Visits concentrate on main thoroughfares (blue to red indicates increasing frequency). (c) Overlap clusters reveal shared route segments; colors identify clusters and stars mark their centers.

3.2 DIVERSITY METRICS

Our diagnostic and evaluation suite combines the structural analysis above with five complementary metrics for assessing the generated corpus. Exit-Direction Entropy (EDE) measures junction-choice balance; Grid Coverage (GC) and Heatmap Spread (HS) describe spatial visitation; Pairwise DTW Distance (PDD) measures route dissimilarity; and Turn-Pattern Diversity (TPD) summarizes local maneuver sequences. Higher values indicate greater diversity for each metric. Implementation details are provided in Appendix E.

Exit-Direction Entropy. We extract junction locations and exits once from the unperturbed scene and baseline trajectories. We match detected sharp turns to the nearest reference junction within a fixed radius and identify the chosen exit from the direction of travel after each turn. For junction i, let $p_{i,k}$ be the fraction of assigned turns taking exit k, K_i the number of exits, and w_i its share of all assigned turns. Then

$$\mathrm{EDE} = \sum_i w_i \frac{-\sum_{k=1}^{K_i} p_{i,k} \log p_{i,k}}{\log K_i}. \tag{1}$$

Grid Coverage and Heatmap Spread. GC is the fraction of traversable grid cells visited by successful trajectories. HS measures how widely visits are distributed: after sorting cell visit counts in descending order, let N_η be the smallest number of cells accounting for a fraction η of all visits. We compute $\mathrm{HS} = N_\eta / N_{\mathrm{free}}$, where N_{free} is the number of traversable cells. GC measures the extent of visitation, while HS distinguishes diffuse visitation from concentration within that extent.

Pairwise DTW Distance. PDD measures trajectory dissimilarity across the evaluated corpus. After resampling each trajectory ξ_i at equal arc-length intervals, we average the Dynamic Time Warping cost (Sakoe & Chiba, 1978) over trajectory pairs, normalized by their mean arc length, where L_i denotes the length of ξ_i:

$$\mathrm{PDD} = \frac{2}{N(N-1)} \sum_{i<j} \frac{2\,\mathrm{DTW}(\xi_i, \xi_j)}{L_i + L_j}. \tag{2}$$

Turn-Pattern Diversity. We discretize heading changes into $\mathcal{A} = \{\mathrm{left}, \mathrm{straight}, \mathrm{right}\}$ and compute the normalized entropy of consecutive token pairs:

$$\mathrm{TPD} = \frac{-\sum_{u,v \in \mathcal{A}} P(u,v) \log P(u,v)}{\log |\mathcal{A}|^2}. \tag{3}$$

This metric complements the spatial measures with the variety of local action sequences.

3.3 QUANTITATIVE DIAGNOSIS

The Base PPO rows in Table 1 show the clearest concentration in spatial visitation. The baseline has the lowest GC among the 25 configurations in the short and long groups, and ranks near the bottom on HS in all three groups. Its EDE is also lowest in the short and medium groups, indicating uneven junction choices. These observations match the concentration on main thoroughfares in Figure 1b.

The remaining metrics describe different aspects of the corpus. Base PPO has the highest pooled PDD in the long group, where routes to different goals can remain far apart while sharing substantial segments. The overlap clusters in Figure 1c reveal this shared structure. TPD changes less across configurations because straight flight dominates the maneuver sequences. We therefore interpret the metrics jointly, using the spatial measures to quantify concentration and PDD and TPD to characterize the resulting route and maneuver distributions.

4 TRAJECTORY-AWARE ENVIRONMENT PERTURBATION

Given an environment E and a corpus $\mathcal{S}$ of successful trajectories, TAEP constructs a virtual obstacle layout $\mathcal{L}_\tau$ from the three diagnostics in Section 3. The generator is then fine-tuned in $E \oplus \mathcal{L}_\tau$ to produce additional successful trajectories. Junction, heatmap, and route-cluster diagnostics identify different candidate locations, allowing the combined intervention to address both local preferences and repeated route segments.

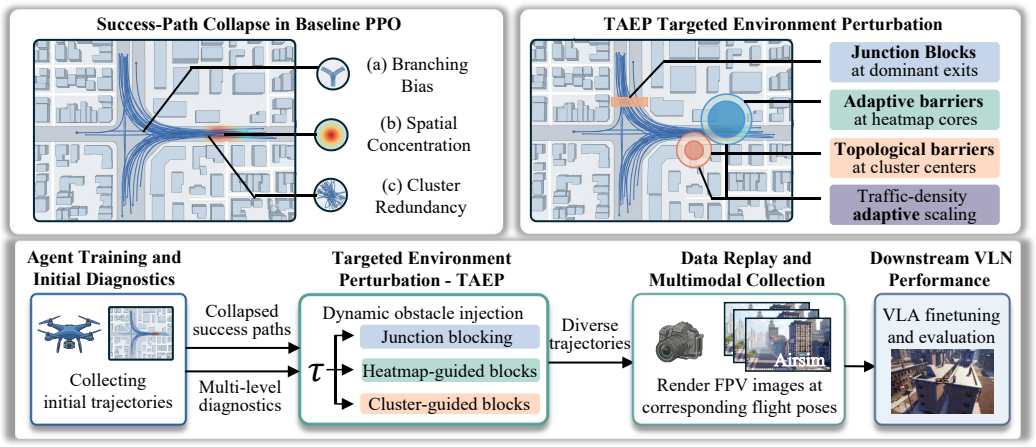


Figure 2: TAEP overview. **Top left.** Successful baseline trajectories reveal branching bias, spatial concentration, and route-cluster redundancy. **Top right.** These diagnostics guide junction blocking and obstacle placement at visitation hotspots and route-cluster centers. **Bottom, left to right.** Train and diagnose the baseline generator; modify the environment at intensity τ and fine-tune the policy to collect diverse trajectories; replay them in AirSim to render first-person observations and aligned actions; fine-tune and evaluate the downstream VLA model.

4.1 Navigation Policy and Two-Stage Training

Policy formulation. The generator uses a shared actor-critic network trained with PPO (Schulman et al., 2017). At step t, its observation is $o_t = (M_t, g_t)$, where M_t is a local occupancy map encoding obstacles and the target, and g_t is the normalized relative goal vector. A CNN encodes M_t, an MLP encodes g_t, and their concatenated features feed the policy and value heads. Actions are sampled as

$$z_t \sim \mathcal{N}(\mu_\theta(o_t), \mathrm{diag}(\sigma_\theta^2)), \qquad a_t = a_{\max} \tanh z_t = (\Delta x_t, \Delta y_t), \tag{4}$$

and the critic $V_\phi(o_t)$ estimates the discounted return. The learning objective is

$$J(\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T-1} \gamma^t r_t\right], \qquad r_t = r_t^{\text{progress}} + r_t^{\text{clearance}} + r_t^{\text{terminal}} + r_t^{\text{step}} + r_t^{\text{milestone}}. \tag{5}$$

Here $r_t^{\text{progress}} = \lambda_g(d_t - d_{t+1})$ rewards reduction in distance to the target building. The remaining terms encourage obstacle clearance, reward arrival or penalize collision, impose a step cost, and provide intermediate rewards at distance milestones to facilitate convergence. Appendix E specifies the reward and training parameters.

Discounting favors earlier receipt of the terminal success reward, making efficient successful routes attractive under this objective. Once a reliable route is learned, discovering other successful routes need not increase task return. This creates a distinction between learning a competent navigator and using it to generate a varied demonstration corpus. TAEP changes the geometry encountered by the policy while retaining the navigation reward and PPO configuration.

Two-stage training. TAEP uses a single diagnosis-and-perturbation sequence (Figure 2). First, the generator is trained in the unperturbed environment and its successful trajectories are collected for diagnosis. Second, virtual obstacles are placed at the diagnosed locations and the policy is fine-tuned in the modified scene. Each stage receives half of the fixed training budget. All perturbation methods start from the same first-stage checkpoint. The selected policy generates trajectories for VLA data construction (Section 4.3; Algorithm A1).

4.2 Diagnosis-Guided Environment Shaping

Intervention strength. We use $\tau \in [0, 1]$ to parameterize intervention strength along a common axis across methods. For environment perturbations, τ controls the obstacle budget. A single module

places up to $K = \mathrm{round}(\tau N_{\max})$ obstacles from its ranked candidates, where $N_{\max}$ is the candidate budget; the full method distributes this budget across modules as described below. For the intrinsic-reward baseline, τ scales the bonus coefficient (Section 5.1). This shared parameterization supports systematic comparison of each method's response to stronger intervention and the trade-off between trajectory diversity and navigation performance.

Perturbation modules. The three modules are entirely data-driven, using successful trajectories and the occupancy map to construct interventions automatically without manual annotation.

(i) *Junction Blocking.* We locate strong turns in the baseline corpus and probe the surrounding occupancy map to identify junctions with multiple exits. Spatial clustering merges nearby detections, and candidates are ranked by occurrence frequency. At selected junctions, a rectangular barrier narrows the dominant exit while retaining a passable gap. This changes the local approach and makes alternative exits more attractive. Clearance checks retract or reject barriers that would seal the passage.

(ii) *Heat-Guided Obstacle Injection.* A visitation heatmap H accumulates the positions along successful trajectories. We select spatially separated high-density cells and scale each circular obstacle by its normalized visitation $\rho_i = H_i / H_{\max}$. A linear interpolation maps this density to a radius between $r_{\min}$ and $r_{\max}$:

$$r_i = (1 - \rho_i) r_{\min} + \rho_i r_{\max}. \tag{6}$$

The bounds are tied to the generator's collision and safety distances, with $r_{\min} = d_{\mathrm{col}}$ and $r_{\max} = d_{\mathrm{safe}} + 2d_{\mathrm{col}}$. Thus the most visited cells receive the largest barriers, while less visited candidates receive smaller obstacles. This concentrates the intervention on dominant traffic regions while keeping obstacle sizes within a controlled range.

(iii) *Corridor Truncation.* We group trajectories by overlap between their visited grid-cell sets G_i, using $|G_i \cap G_j| / \min(|G_i|, |G_j|)$ as the similarity measure. Connected components above an adaptive overlap threshold form route clusters. Circular obstacles are placed at the spatial centers of the dominant clusters, with radii given by Equation 6 after replacing ρ_i with the cluster's trajectory count relative to that of the largest cluster. This targets segments repeatedly used across complete routes. Candidate spacing, clustering, and geometric placement parameters are given in Appendix E.

Module composition. The full method allocates approximately one third of the obstacle budget to each module:

$$K_{\mathrm{heat}} = \left\lceil \frac{\tau N_{\max}}{3} \right\rceil, \quad K_{\mathrm{corr}} = \mathrm{round}\left(\frac{\tau N_{\max}}{3} \right), \quad K_{\mathrm{junc}} = \left\lfloor \frac{\tau N_{\max}}{3} \right\rfloor. \tag{7}$$

Section 5.3 compares the combined and individual modules under matched obstacle budgets.

4.3 From Trajectories to VLN Training Data

Automatic aerial VLN pipelines generate trajectories, render observations, and subsequently synthesize instructions from these sequences (Gao et al., 2025). TAEP supplies the trajectories. Successful paths are smoothed and replayed in AirSim with yaw oriented toward the target to obtain first-person images and synchronized poses. Each demonstration pairs an instruction, the current image, and pose history with the next action $(\Delta x, \Delta y, \Delta\psi)$. The resulting data fine-tune a VLA model with LoRA (Hu et al., 2022) (Appendix C). We will release the generated trajectories and their corresponding first-person-view (FPV) data.

To isolate the effect of trajectory diversity and streamline the experimental analysis, we use the template *"Navigate to the target building at coordinate* (x_g, y_g)*."* This instruction specifies the task independently of the route, giving different successful paths to the same target a consistent task description. Synthesizing a separate description for each path would also change landmark references and directional guidance, coupling behavioral variation with variation in language supervision. The shared template holds task expression constant, allowing the comparison to focus on how trajectory-induced visual, pose, and action coverage affects closed-loop navigation. The same trajectories and rendered images also support richer instruction synthesis in a full VLN data-generation pipeline, which we leave for future work.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Scenes and tasks. We use AirSim16 and AirSim23 from OpenFly (Gao et al., 2025), two photorealistic urban scenes with varied building layouts and interconnected road networks. AirSim16 supports both trajectory-generation and downstream VLA experiments, while AirSim23 provides an additional scene for trajectory-generation evaluation. Occupancy maps are constructed from point-cloud slices at a cruise-altitude band of 10–20 m. Tasks are grouped by Euclidean start-goal distance into short (0–200 m), medium (200–400 m), and long (400–600 m) ranges. AirSim16 includes all three groups; AirSim23 evaluates the long group. For each evaluated scene–distance setting, start-goal pairs are divided into disjoint training (200), validation (200), test (200), and generation (1,230) sets. Training pairs optimize the PPO policy. Validation success rate selects the checkpoint. The selected checkpoint is used both to evaluate trajectory diversity on the test set and to collect successful demonstrations on the generation set for VLA fine-tuning.

Methods and budget. We compare full TAEP and its individual modules with vanilla PPO (Base PPO), count-based exploration (MBIE-EB), and geometry-matched random perturbation (Random). Every run receives 10,000 PPO updates. Perturbation methods use 5,000 clean-environment updates followed by 5,000 updates with obstacles; Base PPO and MBIE-EB train for 10,000 updates in the clean environment. We sweep $\tau \in \{0.2, 0.4, 0.6, 0.8\}$ with $N_{\max} = 12$, giving obstacle budgets of 2, 5, 7, and 10. MBIE-EB adds $\beta/\sqrt{n(s,a)}$ to the task reward, with $\beta = 5\tau$ and counts over position cells and turn bins. Random relocates the shapes used by full TAEP at each intensity to randomly sampled admissible positions. Downstream comparisons also use matched data scales and VLA fine-tuning compute budgets. TAEP's one-time trajectory diagnostics and obstacle placement between training stages incur negligible computational overhead relative to PPO training. Appendix E provides implementation details.

Evaluation. Stochastic rollouts from each selected checkpoint provide the successful trajectories used for the five metrics in Section 3.2. Each setting has 25 configurations: Base PPO and six methods at four intensities. We rank all 25 configurations on each metric and average the five ranks to obtain Mean Rank (lower is better). Table 1 reports the intensity with the lowest Mean Rank for each method; Appendix A reports the complete sweeps. Table 2 averages each metric across the three AirSim16 distance groups and ranks the 16 module–intensity combinations.

5.2 COMPARISON WITH BASELINES

TAEP improves diversity through broader spatial visitation. TAEP achieves the lowest Mean Rank across all three AirSim16 groups, ahead of vanilla PPO, MBIE-EB, and Random (Table 1). Relative to vanilla PPO, short- and long-task GC increase by 35% and 37%, and HS by 23% and 47%, respectively. On medium tasks, the coverage gain is smaller, but TAEP improves EDE and achieves a notable increase in HS. Together, these gains expand coverage and balance visitation, exposing the downstream training corpus to a broader range of navigation situations.

Diagnostic guidance improves perturbation effectiveness. With obstacle shapes and counts matched at each intensity, TAEP achieves a lower selected Mean Rank than Random in every group. On medium tasks, their selected configurations attain nearly identical GC, yet TAEP improves HS by 14.7%. Targeting repeated trajectory structures therefore spreads visits more widely within the covered area, improving the diversity obtained from a fixed obstacle budget.

TAEP expands coverage beyond intrinsic-reward exploration. MBIE-EB improves on vanilla PPO and leads the compared methods in EDE and TPD on long tasks. TAEP, however, exceeds its GC and HS by 19.8% and 22.2%. This suggests that the intrinsic bonus more readily diversifies local decisions and turns, while environment shaping promotes global changes in route coverage. These spatial gains support TAEP's lower Mean Rank in every distance group, with the same overall advantage on AirSim23 (Appendix B).

Table 1: **Main results.** Trajectory diversity on the three distance groups, each method at its selected intensity τ^*. All metrics are higher-is-better (↑); parentheses give the rank among the 25 configurations of that group's full sweep (Appendix A). Best in bold, second best underlined within each block.

Method	τ^*	EDE ↑	GC ↑	HS ↑	PDD ↑	TPD ↑	Mean Rank ↓
Short (0–200 m)							
Base PPO	–	0.7671 (25)	0.2458 (25)	0.0354 (25)	36.27 (3)	0.7386 (9)	17.4
MBIE-EB	0.8	**0.7862** (2)	0.3178 (21)	**0.0438** (5)	35.15 (9)	0.7300 (23)	12.0
Random	0.6	0.7744 (20)	0.3232 (18)	0.0417 (18)	35.53 (5)	0.7313 (22)	16.6
TAEP (full)	0.8	0.7814 (8)	**0.3325** (8)	0.0437 (6)	**37.24** (1)	**0.7418** (7)	**6.0**
Medium (200–400 m)							
Base PPO	–	0.7516 (25)	0.4581 (14)	0.0489 (21)	19.21 (17)	0.6924 (15)	18.4
MBIE-EB	0.4	**0.7714** (4)	0.4462 (18)	0.0512 (14)	19.29 (15)	**0.7008** (4)	11.0
Random	0.8	0.7588 (20)	**0.4795** (3)	0.0470 (23)	**19.73** (11)	0.6969 (10)	13.4
TAEP (full)	0.8	0.7711 (5)	0.4788 (4)	**0.0539** (4)	19.24 (16)	0.6960 (11)	**8.0**
Long (400–600 m)							
Base PPO	–	0.7259 (4)	0.3860 (25)	0.0361 (24)	**14.72** (1)	0.6666 (22)	15.2
MBIE-EB	0.4	**0.7401** (1)	0.4423 (10)	0.0433 (15)	13.91 (8)	**0.7119** (1)	7.0
Random	0.4	0.7118 (9)	0.4569 (7)	0.0500 (5)	13.28 (22)	0.6957 (4)	9.4
TAEP (full)	0.6	0.7117 (10)	**0.5298** (3)	**0.0529** (4)	14.03 (5)	0.6909 (8)	**6.0**

Table 2: **Module ablation.** Metrics averaged over the three distance groups; parentheses give the rank among the 16 ablation entries (4 modules × 4 intensities), and MR averages these five ranks (↓). Per-group sweeps are in Appendix A.

Module	τ	EDE ↑	GC ↑	HS ↑	PDD ↑	TPD ↑	MR ↓
Junction	0.2	0.7440 (13)	0.4217 (5)	0.0456 (14)	23.14 (4)	**0.7164** (1)	7.4
	0.4	0.7460 (10)	0.4175 (6)	0.0453 (15)	22.36 (12)	0.7112 (3)	9.2
	0.6	0.7411 (14)	0.4140 (7)	0.0457 (12)	22.18 (16)	0.7135 (2)	10.2
	0.8	0.7463 (9)	0.4081 (10)	0.0446 (16)	22.90 (6)	0.7079 (7)	9.6
Heat	0.2	0.7441 (12)	0.4044 (12)	0.0462 (6)	22.35 (13)	0.7061 (9)	10.4
	0.4	0.7345 (16)	0.4042 (13)	0.0460 (9)	23.35 (2)	0.7057 (10)	10.0
	0.6	0.7506 (7)	0.4036 (14)	0.0462 (7)	22.66 (8)	0.7052 (11)	9.4
	0.8	0.7370 (15)	0.4004 (16)	0.0457 (11)	23.08 (5)	0.6964 (16)	12.6
Corridor	0.2	0.7457 (11)	0.4026 (15)	0.0457 (12)	22.82 (7)	0.7010 (14)	11.8
	0.4	0.7522 (5)	0.4106 (8)	0.0458 (10)	22.65 (9)	0.7083 (5)	7.4
	0.6	0.7511 (6)	0.4103 (9)	0.0462 (7)	22.33 (14)	0.7104 (4)	8.0
	0.8	0.7491 (8)	0.4065 (11)	0.0467 (5)	22.61 (10)	0.7038 (12)	9.2
Full TAEP	0.2	**0.7614** (1)	0.4542 (2)	0.0497 (4)	22.32 (15)	0.7036 (13)	7.0
	0.4	0.7613 (2)	**0.4594** (1)	0.0503 (2)	22.46 (11)	0.7004 (15)	6.2
	0.6	0.7528 (4)	0.4446 (4)	0.0498 (3)	23.21 (3)	0.7076 (8)	4.4
	0.8	0.7593 (3)	0.4530 (3)	**0.0508** (1)	**23.41** (1)	0.7081 (6)	**2.8**

5.3 Ablation Study

Combining diagnostics improves multiple aspects of diversity. Across the four tested intensities, full TAEP achieves higher distance-averaged EDE, GC, and HS than the individual modules (Table 2). Distributing the budget across candidate types thus improves junction balance and spatial visitation together. Individual modules show distinct strengths—Junction has the highest TPD, while Heat achieves high PDD at $\tau = 0.4$—with effects spanning several metrics beyond their primary diagnostic targets. This supports their use as complementary placement criteria.

Additional obstacles are most effective when distributed across structures. Full TAEP's Mean Rank improves from 7.0 to 2.8 as intensity increases, while single modules peak earlier, at $\tau = 0.2$, 0.6, and 0.4 for Junction, Heat, and Corridor. The full method maintains strong spatial metrics

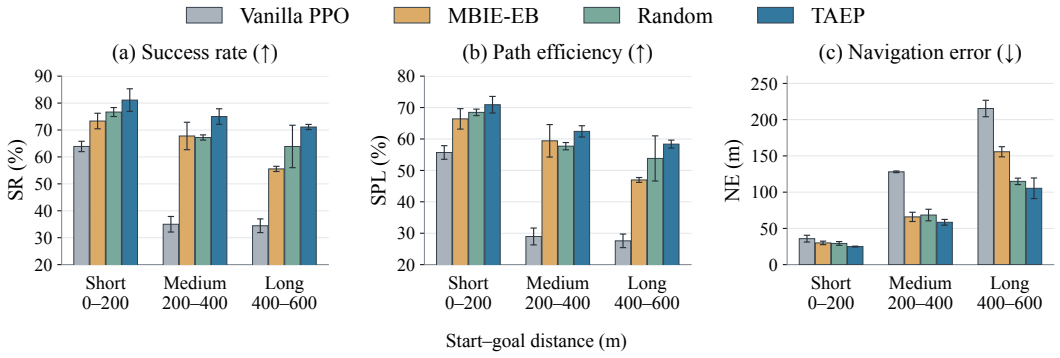


Figure 3: Downstream VLA navigation with demonstrations from four trajectory generators. Panels show SR (↑), SPL (↑), and NE (↓) across the three distance groups. Bars show means over three evaluation rounds; error bars indicate sample standard deviations.

throughout and achieves the highest HS and PDD at $\tau = 0.8$. Under matched obstacle budgets, these results favor distributing interventions across diagnosed structures. Per-group sweeps reinforce this finding on long tasks, where hotspot perturbation alone becomes less effective at higher perturbation intensities, while full TAEP retains a substantial coverage advantage (Appendix A).

5.4 Downstream VLA Navigation

We fine-tune the same VLA backbone on demonstrations from vanilla PPO, MBIE-EB, Random, and TAEP for eight epochs under a common instruction format and training protocol (Section 4.3; Appendix C). Each corpus contains approximately 2,500 training trajectories and 170,000 frames. VLA evaluation uses a fixed set of 180 randomly generated start-goal pairs, 60 per distance group, shared across all methods and evaluation rounds. Figure 3 and Table A6 summarize three evaluation rounds, with equal weighting across groups.

TAEP leads all three baselines on every navigation metric in each distance group, raising mean SR from 44.4% with vanilla PPO data to 75.7%. The benefits are greatest on medium and long tasks; on the longest tasks, TAEP more than doubles vanilla PPO's SR and roughly halves NE. These gains highlight the value of broader demonstration coverage over extended navigation horizons.

Even random environment perturbation outperforms MBIE-EB on all three aggregate metrics, reinforcing the advantage of changing traversal conditions over the auxiliary exploration reward for this data-generation task. TAEP adds a further 6.5 percentage points in mean SR over the strongest baseline, with larger gains on medium and long tasks. Thus diagnostic placement strengthens the downstream benefits of environment shaping beyond those obtained from unguided perturbation.

6 Conclusion

Scalable training data for aerial VLN requires flight demonstrations that cover a range of navigation behavior. We study Success-Path Collapse, in which a successful trajectory generator repeatedly uses similar directions, locations, and route segments. TAEP uses these patterns to guide virtual obstacle placement and fine-tunes the generator to discover alternative routes. Across the evaluated settings, TAEP achieves better aggregate trajectory diversity than vanilla PPO, count-based intrinsic rewards, and geometry-matched random perturbation. Averaged across distance groups, TAEP demonstrations improve downstream VLA success rate by 31.3 percentage points over vanilla PPO and 6.5 over the strongest baseline, while reducing navigation error by 50.3% relative to vanilla PPO. By linking corpus diagnostics to environment edits, TAEP makes behavioral coverage an explicit consideration in aerial VLN data generation. Future work can extend TAEP into an adaptive curriculum by repeatedly diagnosing newly collected trajectories and updating the environment as the policy learns. Richer interventions, such as moving obstacles and dynamic agents, could further diversify the interactions and avoidance maneuvers represented in aerial VLN demonstrations.

AI USE STATEMENT

Generative AI tools were used to polish the manuscript and assist with code writing. The authors reviewed all AI-assisted text and code and take responsibility for the final work.

REFERENCES

P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. van den Hengel. Vision-and-Language Navigation: Interpreting Visually-Grounded Navigation Instructions in Real Environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3674–3683, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Anderson_Vision-and-Language_Navigation_Interpreting_CVPR_2018_paper.html.

M. Bellemare, S. Srinivasan, G. Ostrovski, T. Schaul, D. Saxton, and R. Munos. Unifying Count-Based Exploration and Intrinsic Motivation. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/hash/afda332245e2af431fb7b672a68b659d-Abstract.html.

Y. Burda, H. Edwards, A. Storkey, and O. Klimov. Exploration by Random Network Distillation. arXiv preprint arXiv:1810.12894, 2018. URL https://arxiv.org/abs/1810.12894.

H. Chen, A. Suhr, D. Misra, N. Snavely, and Y. Artzi. TOUCHDOWN: Natural Language Navigation and Spatial Reasoning in Visual Street Environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12538–12547, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Chen_TOUCHDOWN_Natural_Language_Navigation_and_Spatial_Reasoning_in_Visual_Street_CVPR_2019_paper.html.

M. Dennis, N. Jaques, E. Vinitsky, A. Bayen, S. Russell, A. Critch, and S. Levine. Emergent Complexity and Zero-shot Transfer via Unsupervised Environment Design. In *Advances in Neural Information Processing Systems*, volume 33, pp. 13049–13061. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/985e9a46e10005356bbaf194249f6856-Abstract.html.

B. Eysenbach, A. Gupta, J. Ibarz, and S. Levine. Diversity is All You Need: Learning Skills without a Reward Function. arXiv preprint arXiv:1802.06070, 2018. URL https://arxiv.org/abs/1802.06070.

D. Fried, R. Hu, V. Cirik, A. Rohrbach, J. Andreas, L.-P. Morency, T. Berg-Kirkpatrick, K. Saenko, D. Klein, and T. Darrell. Speaker-Follower Models for Vision-and-Language Navigation. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/hash/6a81681a7af700c6385d36577ebec359-Abstract.html.

Y. Gao, C. Li, Z. You, J. Liu, Z. Li, P. Chen, Q. Chen, Z. Tang, L. Wang, P. Yang, et al. OpenFly: A Versatile Toolchain and Large-scale Benchmark for Aerial Vision-Language Navigation. arXiv preprint arXiv:2502.18041, 2025. URL https://arxiv.org/abs/2502.18041v1.

P. Hart, N. Nilsson, and B. Raphael. A Formal Basis for the Heuristic Determination of Minimum Cost Paths. *IEEE Transactions on Systems Science and Cybernetics*, 4(2):100–107, 1968. doi: 10.1109/tssc.1968.300136. URL https://doi.org/10.1109/tssc.1968.300136.

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, et al. OpenVLA: An Open-Source Vision-Language-Action Model. arXiv preprint arXiv:2406.09246, 2024. URL https://arxiv.org/abs/2406.09246.

J. Krantz, E. Wijmans, A. Majumdar, D. Batra, and S. Lee. Beyond the Nav-Graph: Vision-and-Language Navigation in Continuous Environments. In *European Conference on Computer Vision (ECCV) 2020*, pp. 104–120. Springer, 2020. doi: 10.1007/978-3-030-58604-1_7. URL https://doi.org/10.1007/978-3-030-58604-1_7.

J. Li, H. Tan, and M. Bansal. EnvEdit: Environment Editing for Vision-and-Language Navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15407–15417, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Li_EnvEdit_Environment_Editing_for_Vision-and-Language_Navigation_CVPR_2022_paper.html.

J. Li, X. Huang, Z. Wang, Y. Wang, W. Ge, Q. He, M. Zhu, F. Gao, Y. Wu, and X. Zhou. FlyMirage: A Fully Automated Generation Pipeline for Diverse and Scalable UAV Flight Data via Generative World Model. arXiv preprint arXiv:2605.19600, 2026. URL https://arxiv.org/abs/2605.19600.

C. Liu, F. Zhu, X. Chang, X. Liang, Z. Ge, and Y.-D. Shen. Vision-Language Navigation With Random Environmental Mixup. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1644–1654, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Liu_Vision-Language_Navigation_With_Random_Environmental_Mixup_ICCV_2021_paper.html.

S. Liu, H. Zhang, Y. Qi, P. Wang, Y. Zhang, and Q. Wu. AerialVLN: Vision-and-Language Navigation for UAVs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15384–15394, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Liu_AerialVLN_Vision-and-Language_Navigation_for_UAVs_ICCV_2023_paper.html.

J. Parker-Holder, M. Jiang, M. Dennis, M. Samvelyan, J. Foerster, E. Grefenstette, and T. Rocktäschel. Evolving Curricula with Regret-Based Environment Design. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 17473–17498. PMLR, 2022. URL https://proceedings.mlr.press/v162/parker-holder22a.html.

D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell. Curiosity-driven Exploration by Self-supervised Prediction. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2778–2787. PMLR, 2017. URL https://proceedings.mlr.press/v70/pathak17a.html.

S. Ross, G. Gordon, and D. Bagnell. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635. PMLR, 2011. URL https://proceedings.mlr.press/v15/ross11a.html.

H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978. doi: 10.1109/tassp.1978.1163055. URL https://doi.org/10.1109/tassp.1978.1163055.

J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal Policy Optimization Algorithms. arXiv preprint arXiv:1707.06347, 2017. URL https://arxiv.org/abs/1707.06347.

A. L. Strehl and M. L. Littman. An analysis of model-based Interval Estimation for Markov Decision Processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008. doi: 10.1016/j.jcss.2007.08.009. URL https://doi.org/10.1016/j.jcss.2007.08.009.

H. Tan, L. Yu, and M. Bansal. Learning to Navigate Unseen Environments: Back Translation with Environmental Dropout. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2610–2621. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1268. URL https://aclanthology.org/N19-1268.bib.

R. Wang, J. Lehman, J. Clune, and K. O. Stanley. Paired Open-Ended Trailblazer (POET): Endlessly Generating Increasingly Complex and Diverse Learning Environments and Their Solutions. arXiv preprint arXiv:1901.01753, 2019. URL https://arxiv.org/abs/1901.01753.

X. Wang, D. Yang, Y. Liao, W. Zheng, W. Wu, B. Dai, H. Li, and S. Liu. UAV-Flow Colosseo: A Real-World Benchmark for Flying-on-a-Word UAV Imitation Learning. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025a. doi: 10.52202/085713-3391. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/92cfa104db7fdd053492df3589ddfd49-Abstract-Datasets_and_Benchmarks_Track.html.

X. Wang, D. Yang, Z. Wang, H. Kwan, J. Chen, W. Wu, H. Li, Y. Liao, and S. Liu. Towards Realistic UAV Vision-Language Navigation: Platform, Benchmark, and Methodology. In *International Conference on Learning Representations (ICLR)*, 2025b. URL https://openreview.net/forum?id=rUvCIvI4eB.

Z. Wang, J. Li, Y. Hong, Y. Wang, Q. Wu, M. Bansal, S. Gould, H. Tan, and Y. Qiao. Scaling Data Generation in Vision-and-Language Navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12009–12020, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Wang_Scaling_Data_Generation_in_Vision-and-Language_Navigation_ICCV_2023_paper.html.

Y. Zhong, R. Zhang, Z. Zhang, S. Wang, C. Fang, X. Zhang, J. Guo, S. Peng, D. Huang, Y. Yan, et al. World-Consistent Data Generation for Vision-and-Language Navigation. arXiv preprint arXiv:2412.06413, 2024. URL https://arxiv.org/abs/2412.06413.

B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid, et al. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. In *Proceedings of the Conference on Robot Learning (CoRL)*, volume 229 of *Proceedings of Machine Learning Research*, pp. 2165–2183. PMLR, 2023. URL https://proceedings.mlr.press/v229/zitkovich23a.html.

APPENDIX

Appendix A reports the complete AirSim16 diversity sweeps. Appendix B presents the AirSim23 results. Appendix C describes VLA training and evaluation, Appendix D gives the TAEP pseudocode, and Appendix E specifies the generator, metrics, and obstacle-placement parameters.

A FULL DIVERSITY RESULTS

Tables A1–A3 report all 25 configurations for each AirSim16 distance group. Each cell contains a metric value and its rank within that group; Mean Rank averages the five ranks. The main comparison selects each method's lowest Mean Rank, while the module ablation averages metric values across the three groups before ranking configurations.

On short tasks, vanilla PPO ranks last on EDE, GC, and HS, and full TAEP at $\tau = 0.8$ has the best Mean Rank. Medium tasks show more differentiated responses: Heat achieves the highest HS, Corridor the highest PDD, and Junction the highest GC. Full TAEP combines strong EDE, GC, and HS to obtain the best Mean Rank. On long tasks, MBIE-EB is competitive, particularly on EDE and TPD, while full TAEP achieves the highest GC and HS values across the sweep. The best full-TAEP intensity is 0.6 by Mean Rank and 0.8 by the two spatial metrics, illustrating the trade-offs summarized by the aggregate ranking.

Table A1: Complete intensity sweep on the short group (0–200 m Euclidean start-goal distance). All metrics are computed over all successfully generated trajectories and are higher-is-better (↑); each cell gives the value with its rank among the 25 configurations of the group. Mean Rank averages the five per-metric ranks (↓). Best and second-best values among the 25 configurations are bold and underlined, respectively; ties share the same marking.

Method	τ	EDE↑	GC↑	HS↑	PDD↑	TPD↑	Mean Rank↓
Base PPO	–	0.7671 (25)	0.2458 (25)	0.0354 (25)	36.27 (3)	0.7386 (9)	17.4
MBIE-EB	0.2	**0.7882** (1)	0.3089 (23)	0.0413 (21)	34.57 (17)	0.7442 (4)	13.2
	0.4	0.7854 (3)	0.2996 (24)	0.0412 (22)	34.29 (19)	0.7358 (16)	16.8
	0.6	0.7730 (22)	0.3091 (22)	0.0411 (23)	34.05 (23)	0.7239 (25)	23.0
	0.8	<u>0.7862</u> (2)	0.3178 (21)	0.0438 (5)	35.15 (9)	0.7300 (23)	12.0
Random	0.2	0.7754 (18)	0.3192 (19)	0.0410 (24)	33.76 (24)	0.7422 (6)	18.2
	0.4	0.7738 (21)	0.3302 (11)	0.0413 (20)	34.94 (12)	0.7340 (19)	16.6
	0.6	0.7744 (20)	0.3232 (18)	0.0417 (18)	35.53 (5)	0.7313 (22)	16.6
	0.8	0.7683 (24)	0.3186 (20)	0.0417 (19)	34.80 (16)	0.7326 (21)	20.0
TAEP-Junction	0.2	0.7822 (6)	0.3262 (17)	0.0419 (16)	35.45 (6)	0.7386 (10)	11.0
	0.4	0.7823 (5)	0.3460 (3)	0.0434 (9)	34.17 (20)	0.7377 (11)	9.6
	0.6	0.7829 (4)	0.3266 (15)	0.0435 (8)	35.11 (10)	0.7436 (5)	8.4
	0.8	0.7809 (9)	0.3264 (16)	0.0427 (13)	34.84 (15)	0.7395 (8)	12.2
TAEP-Heat	0.2	0.7788 (12)	0.3278 (13)	0.0424 (14)	34.14 (22)	<u>0.7451</u> (2)	12.6
	0.4	0.7790 (11)	0.3296 (12)	0.0431 (11)	<u>36.71</u> (2)	0.7376 (12)	9.6
	0.6	0.7785 (13)	0.3317 (9)	0.0443 (4)	34.87 (13)	**0.7464** (1)	8.0
	0.8	0.7796 (10)	0.3375 (6)	<u>0.0461</u> (2)	35.85 (4)	0.7361 (15)	<u>7.4</u>
TAEP-Corridor	0.2	0.7785 (14)	0.3350 (7)	0.0419 (17)	34.47 (18)	0.7363 (14)	14.0
	0.4	0.7727 (23)	0.3277 (14)	0.0422 (15)	34.86 (14)	0.7351 (18)	16.8
	0.6	0.7759 (17)	0.3314 (10)	0.0433 (10)	35.01 (11)	0.7442 (3)	10.2
	0.8	0.7759 (16)	0.3444 (4)	0.0431 (12)	33.07 (25)	0.7354 (17)	14.8
TAEP-Combined	0.2	0.7771 (15)	<u>0.3640</u> (2)	0.0459 (3)	34.17 (21)	0.7333 (20)	12.2
	0.4	0.7746 (19)	**0.3653** (1)	**0.0467** (1)	35.17 (8)	0.7280 (24)	10.6
	0.6	0.7815 (7)	0.3376 (5)	0.0436 (7)	35.37 (7)	0.7371 (13)	7.8
	0.8	0.7814 (8)	0.3325 (8)	0.0437 (6)	**37.24** (1)	0.7418 (7)	**6.0**

Table A2: Complete intensity sweep on the medium group (200–400 m). Format follows Table A1.

Method	τ	EDE↑	GC↑	HS↑	PDD↑	TPD↑	Mean Rank↓
Base PPO	–	0.7516 (25)	0.4581 (14)	0.0489 (21)	19.21 (17)	0.6924 (15)	18.4
MBIE-EB	0.2	0.7697 (8)	0.4420 (19)	0.0524 (7)	19.12 (18)	0.6864 (24)	15.2
	0.4	0.7714 (4)	0.4462 (18)	0.0512 (14)	19.29 (15)	0.7008 (4)	11.0
	0.6	**0.7828** (1)	0.4410 (21)	0.0494 (20)	18.72 (23)	<u>0.7062</u> (2)	13.4
	0.8	0.7558 (24)	0.4479 (17)	0.0535 (5)	18.87 (21)	0.6792 (25)	18.4
Random	0.2	0.7579 (21)	0.4016 (24)	0.0463 (24)	19.81 (9)	0.6920 (16)	18.8
	0.4	0.7692 (10)	0.4375 (23)	0.0474 (22)	19.53 (13)	0.6954 (12)	16.0
	0.6	0.7579 (22)	0.3880 (25)	0.0443 (25)	19.90 (7)	**0.7064** (1)	16.0
	0.8	0.7588 (20)	0.4795 (3)	0.0470 (23)	19.73 (11)	0.6969 (10)	13.4
TAEP-Junction	0.2	0.7636 (15)	**0.5017** (1)	0.0495 (19)	20.12 (4)	0.6987 (6)	9.0
	0.4	0.7709 (6)	0.4676 (9)	0.0512 (15)	19.85 (8)	0.7058 (3)	<u>8.2</u>
	0.6	<u>0.7764</u> (2)	0.4743 (6)	0.0512 (16)	17.98 (25)	0.6980 (7)	11.2
	0.8	0.7695 (9)	0.4752 (5)	0.0495 (18)	19.74 (10)	0.6911 (21)	12.6
TAEP-Heat	0.2	0.7692 (11)	0.4419 (20)	0.0521 (9)	18.91 (20)	0.6919 (17)	15.4
	0.4	0.7573 (23)	0.4536 (15)	0.0543 (3)	19.94 (6)	0.6989 (5)	10.4
	0.6	0.7618 (18)	0.4485 (16)	<u>0.0546</u> (2)	19.95 (5)	0.6914 (18)	11.8
	0.8	0.7618 (19)	0.4407 (22)	**0.0558** (1)	<u>20.29</u> (2)	0.6950 (13)	11.4
TAEP-Corridor	0.2	0.7699 (7)	0.4650 (12)	0.0518 (10)	19.73 (12)	0.6901 (23)	12.8
	0.4	0.7689 (13)	0.4663 (11)	0.0524 (8)	19.03 (19)	0.6909 (22)	14.6
	0.6	0.7624 (17)	0.4710 (7)	0.0515 (11)	18.29 (24)	0.6912 (19)	15.6
	0.8	0.7625 (16)	0.4600 (13)	0.0514 (12)	**21.18** (1)	0.6980 (8)	10.0
TAEP-Combined	0.2	0.7691 (12)	0.4679 (8)	0.0497 (17)	19.31 (14)	0.6975 (9)	12.0
	0.4	0.7752 (3)	<u>0.4840</u> (2)	0.0513 (13)	18.79 (22)	0.6911 (20)	12.0
	0.6	0.7652 (14)	0.4665 (10)	0.0528 (6)	20.22 (3)	0.6947 (14)	9.4
	0.8	0.7711 (5)	0.4788 (4)	0.0539 (4)	19.24 (16)	0.6960 (11)	**8.0**

Table A3: Complete intensity sweep on the long group (400–600 m). Format follows Table A1.

Method	τ	EDE↑	GC↑	HS↑	PDD↑	TPD↑	Mean Rank↓
Base PPO	–	0.7259 (4)	0.3860 (25)	0.0361 (24)	**14.72** (1)	0.6666 (22)	15.2
MBIE-EB	0.2	0.6768 (21)	0.4537 (8)	0.0441 (12)	13.36 (21)	0.6902 (10)	14.4
	0.4	**0.7401** (1)	0.4423 (10)	0.0433 (15)	13.91 (8)	**0.7119** (1)	7.0
	0.6	0.7249 (6)	0.4673 (5)	0.0493 (6)	13.78 (11)	0.6923 (7)	7.0
	0.8	0.6978 (13)	0.4579 (6)	0.0493 (7)	13.63 (14)	0.6944 (5)	9.0
Random	0.2	0.6764 (22)	0.4151 (21)	0.0482 (8)	13.51 (16)	0.6798 (18)	17.0
	0.4	0.7118 (9)	0.4569 (7)	0.0500 (5)	13.28 (22)	0.6957 (4)	9.4
	0.6	0.6867 (17)	0.4399 (12)	0.0419 (19)	13.97 (7)	0.6871 (12)	13.4
	0.8	0.6872 (16)	0.4072 (24)	0.0440 (13)	13.81 (10)	0.6906 (9)	14.4
TAEP-Junction	0.2	0.6862 (18)	0.4373 (15)	0.0454 (10)	13.85 (9)	**0.7119** (1)	10.6
	0.4	0.6849 (19)	0.4390 (13)	0.0414 (21)	13.07 (25)	0.6901 (11)	17.8
	0.6	0.6639 (25)	0.4411 (11)	0.0423 (18)	13.44 (18)	0.6988 (2)	14.8
	0.8	0.6886 (15)	0.4227 (20)	0.0415 (20)	14.12 (3)	0.6930 (6)	12.8
TAEP-Heat	0.2	0.6843 (20)	0.4435 (9)	0.0442 (11)	14.01 (6)	0.6812 (15)	12.2
	0.4	0.6673 (24)	0.4293 (17)	0.0406 (22)	13.39 (20)	0.6805 (16)	19.8
	0.6	0.7114 (11)	0.4306 (16)	0.0397 (23)	13.16 (23)	0.6777 (20)	18.6
	0.8	0.6696 (23)	0.4230 (19)	0.0353 (25)	13.09 (24)	0.6582 (23)	22.8
TAEP-Corridor	0.2	0.6886 (14)	0.4078 (23)	0.0433 (16)	14.25 (2)	0.6766 (21)	15.2
	0.4	0.7151 (7)	0.4377 (14)	0.0427 (17)	14.06 (4)	0.6988 (2)	8.8
	0.6	0.7149 (8)	0.4284 (18)	0.0438 (14)	13.69 (13)	0.6958 (3)	11.2
	0.8	0.7088 (12)	0.4150 (22)	0.0456 (9)	13.59 (15)	0.6780 (19)	15.4
TAEP-Combined	0.2	0.7379 (2)	0.5307 (2)	0.0534 (2)	13.48 (17)	0.6799 (17)	8.0
	0.4	0.7342 (3)	0.5289 (4)	0.0529 (3)	13.43 (19)	0.6821 (14)	8.6
	0.6	0.7117 (10)	0.5298 (3)	0.0529 (4)	14.03 (5)	0.6909 (8)	**6.0**
	0.8	0.7254 (5)	**0.5478** (1)	**0.0548** (1)	13.74 (12)	0.6864 (13)	6.4

B ADDITIONAL RESULTS ON AIRSIM23

We evaluate the long-distance group (400–600 m) in AirSim23 using the same trajectory-generation methods, training budgets, and diversity metrics. Table A4 compares full TAEP with vanilla PPO, MBIE-EB, and Random at their selected intensities. TAEP has the best Mean Rank (8.2), followed by MBIE-EB (11.0), Random (11.8), and vanilla PPO (12.8). MBIE-EB has the highest EDE and GC among these methods, while TAEP has the highest HS and PDD. Vanilla PPO has the highest TPD. The aggregate advantage of TAEP here is driven by spatial spread and route dissimilarity, with a different metric profile from AirSim16.

The module sweep in Table A5 also favors the full method at $\tau = 0.8$. Its Mean Rank improves across the tested intensities, whereas individual modules peak at intermediate intensities. Junction is the strongest single-module configuration by Mean Rank. At the metric level, Heat reaches the highest PDD and nearly the highest HS, while Junction preserves more turn-pattern diversity. These results reinforce the value of combining placement criteria while showing that their individual effects depend on scene geometry.

Table A4: AirSim23, long group (400–600 m): comparison. Each perturbation method is shown at its best intensity τ^*. Parentheses give the rank among the 25 configurations of the sweep. Best and second-best values among the four displayed methods are bold and underlined, respectively.

Method	τ^*	EDE↑	GC↑	HS↑	PDD↑	TPD↑	Mean Rank↓
Base PPO	–	0.7624 (18)	0.6912 (21)	<u>0.1237</u> (9)	20.47 (15)	**0.5692** (1)	12.8
MBIE-EB	0.6	**0.7844** (2)	**0.7354** (1)	0.1220 (11)	19.95 (25)	0.5327 (16)	<u>11.0</u>
Random	0.4	0.7651 (11)	0.7064 (8)	0.1200 (21)	<u>20.48</u> (13)	<u>0.5417</u> (5)	11.8
TAEP (full)	0.8	<u>0.7743</u> (5)	<u>0.7070</u> (7)	**0.1381** (1)	**21.48** (3)	0.5202 (25)	**8.2**

Table A5: AirSim23, long group: module ablation. Ranks are again taken over all 25 configurations of the sweep. Best and second-best values among the 16 displayed module–intensity configurations are bold and underlined, respectively.

Method	τ	EDE↑	GC↑	HS↑	PDD↑	TPD↑	Mean Rank↓
TAEP-Junction	0.2	0.7651 (12)	0.6946 (18)	0.1208 (18)	20.88 (8)	**0.5433** (2)	11.6
	0.4	0.7627 (16)	0.6900 (22)	0.1211 (15)	20.83 (11)	0.5339 (15)	15.8
	0.6	**0.7770** (3)	0.6936 (19)	0.1259 (7)	20.87 (9)	0.5384 (10)	<u>9.6</u>
	0.8	0.7728 (7)	0.6777 (25)	0.1254 (8)	<u>21.81</u> (2)	0.5370 (14)	11.2
TAEP-Heat	0.2	0.7610 (20)	0.6959 (17)	0.1300 (6)	21.15 (5)	0.5257 (19)	13.4
	0.4	<u>0.7754</u> (4)	0.6874 (23)	<u>0.1335</u> (2)	21.03 (6)	0.5210 (23)	11.6
	0.6	0.7741 (6)	0.6929 (20)	0.1327 (3)	21.31 (4)	0.5234 (22)	11.0
	0.8	0.7708 (8)	0.6816 (24)	0.1307 (5)	**21.90** (1)	0.5206 (24)	12.4
TAEP-Corridor	0.2	0.7628 (15)	0.7064 (8)	0.1192 (22)	20.23 (21)	0.5388 (8)	14.8
	0.4	0.7608 (21)	0.7054 (12)	0.1219 (12)	20.92 (7)	<u>0.5430</u> (4)	11.2
	0.6	0.7518 (25)	0.7014 (14)	0.1215 (13)	20.06 (24)	0.5392 (7)	16.6
	0.8	0.7625 (17)	0.7011 (15)	0.1207 (19)	20.28 (18)	0.5401 (6)	15.0
TAEP-Combined	0.2	0.7526 (24)	<u>0.7098</u> (6)	0.1210 (16)	20.12 (23)	0.5376 (12)	16.2
	0.4	0.7569 (22)	0.7062 (10)	0.1319 (4)	20.26 (20)	0.5284 (18)	14.8
	0.6	0.7612 (19)	**0.7150** (4)	0.1223 (10)	20.86 (10)	0.5378 (11)	10.8
	0.8	0.7743 (5)	0.7070 (7)	**0.1381** (1)	21.48 (3)	0.5202 (25)	**8.2**

C VLA Training and Evaluation Details

Data construction. Successful trajectories are replayed in AirSim at a fixed altitude of 15 m, with yaw oriented toward the target, to obtain first-person RGB images and synchronized poses. Each example contains the instruction *"Navigate to the target building at coordinate* (x_g, y_g)*."*, the current image and pose, the target coordinate, and the previous 20 poses. The prediction target is the next action $(\Delta x, \Delta y, \Delta\psi)$. Each action dimension is discretized into 256 bins between its 1st and 99th percentiles in the corresponding corpus. Bins are mapped to reserved language-model tokens. All corpora use the same processing procedure.

Fine-tuning. We fine-tune Qwen2.5-VL-7B-Instruct with LoRA (Hu et al., 2022), using rank 32, scaling factor 64, and dropout 0.05. LoRA is applied to all linear layers; embeddings and the output head are also trainable. Supervision is applied only to the assistant response, which consists of three action tokens encoding $(\Delta x, \Delta y, \Delta\psi)$ followed by an end-of-response token. All VLA models are trained for eight epochs using AdamW with a learning rate of 2×10^{-5} after a 3% warm-up, batch size 14 per device on four GPUs, ZeRO-2, bf16 precision, and gradient checkpointing.

Closed-loop evaluation. The model receives the current image and pose history, predicts three action tokens using greedy decoding, and executes the decoded displacement before receiving the next observation. Decoding uses the action-percentile file of the corresponding corpus. We randomly generate 60 start-goal pairs per distance group once, yielding a fixed set of 180 pairs shared across all methods and all three evaluation rounds. An episode succeeds when the UAV reaches within 20 m of the target building in at most 200 action steps without collision.

Navigation metrics. For N evaluated episodes, let $S_i \in \{0, 1\}$ indicate success, p_i be the executed path length, and d_i the final distance to the target building. The reference length ℓ_i is the shortest collision-free path between the start and goal, computed using A* (Hart et al., 1968). We report

$$\mathrm{SR} = \frac{1}{N}\sum_{i=1}^{N} S_i, \qquad \mathrm{SPL} = \frac{1}{N}\sum_{i=1}^{N} S_i \frac{\ell_i}{\max(\ell_i, p_i)}, \qquad \mathrm{NE} = \frac{1}{N}\sum_{i=1}^{N} d_i. \tag{8}$$

SR and SPL are reported as percentages, and NE in meters.

Result aggregation. Table A6 reports means and sample standard deviations across the three completed evaluation rounds for each method. SR and SPL are percentages; NE is in meters. For the macro-average, each metric is first averaged equally over the three distance groups within each round, then summarized across rounds.

Table A6: Downstream VLA navigation. Values are mean ± sample standard deviation over three evaluation rounds. Macro-averages give equal weight to the three distance groups. Best and second-best means within each block are bold and underlined, respectively.

Distance	Generator	SR (%) ↑	SPL (%) ↑	NE (m) ↓
0–200 m	Vanilla PPO	63.89 ± 1.92	55.70 ± 2.17	35.85 ± 4.68
	MBIE-EB	73.33 ± 2.89	66.40 ± 3.25	29.99 ± 2.42
	Random	$\underline{76.67} \pm 1.67$	$\underline{68.47} \pm 1.03$	$\underline{29.26} \pm 2.58$
	TAEP (full)	$\mathbf{81.11} \pm 4.19$	$\mathbf{70.91} \pm 2.64$	$\mathbf{24.92} \pm 0.59$
200–400 m	Vanilla PPO	35.00 ± 2.89	28.97 ± 2.68	128.04 ± 1.16
	MBIE-EB	$\underline{67.78} \pm 5.09$	$\underline{59.41} \pm 5.17$	$\underline{65.89} \pm 6.37$
	Random	67.22 ± 0.96	57.71 ± 1.16	68.42 ± 7.97
	TAEP (full)	$\mathbf{75.00} \pm 2.89$	$\mathbf{62.42} \pm 1.79$	$\mathbf{58.44} \pm 3.94$
400–600 m	Vanilla PPO	34.44 ± 2.55	27.58 ± 2.16	215.38 ± 11.34
	MBIE-EB	55.56 ± 0.96	46.96 ± 0.78	155.72 ± 6.99
	Random	$\underline{63.89} \pm 7.88$	$\underline{53.81} \pm 7.19$	$\underline{114.86} \pm 4.43$
	TAEP (full)	$\mathbf{71.11} \pm 0.96$	$\mathbf{58.36} \pm 1.27$	$\mathbf{105.31} \pm 14.22$
Macro-average	Vanilla PPO	44.44 ± 2.22	37.42 ± 1.81	126.42 ± 5.03
	MBIE-EB	65.56 ± 2.78	57.59 ± 2.64	83.87 ± 4.64
	Random	$\underline{69.26} \pm 2.31$	$\underline{60.00} \pm 2.61$	$\underline{70.85} \pm 1.80$
	TAEP (full)	$\mathbf{75.74} \pm 2.51$	$\mathbf{63.90} \pm 1.62$	$\mathbf{62.89} \pm 3.57$

D TAEP PIPELINE PSEUDOCODE

Algorithm A1 summarizes the two-stage generator training and subsequent use of its selected checkpoint. The first-stage checkpoint is shared by the perturbation methods. The training, validation, test, and generation task sets have distinct roles as specified in Section 5.1.

Algorithm A1 TAEP data-generation pipeline

Require: Map E, task sets $T_{\text{train}}, T_{\text{val}}, T_{\text{test}}, T_{\text{gen}}$, intensity τ
Ensure: Diversity measurements m, demonstration corpus $\mathcal{D}$
1: Train π_0 in E on T_{train} for U_1 PPO updates
2: Collect successful training trajectories $\mathcal{S}$ from π_0
3: Diagnose junctions, visitation heat, and route clusters from $\mathcal{S}$
4: Construct obstacle layout $\mathcal{L}_\tau$ from these diagnostics
5: Fine-tune π_0 in $E \oplus \mathcal{L}_\tau$ for U_2 PPO updates on T_{train}
6: Select checkpoint π^* by success rate on T_{val}
7: $m \leftarrow$ diversity metrics from successful rollouts of π^* on T_{test}
8: $\mathcal{D} \leftarrow$ successful stochastic rollouts of π^* on T_{gen}
9: **return** $m, \mathcal{D}$

E IMPLEMENTATION DETAILS

Generator and optimization. The occupancy input has two channels and a 100×100 grid covering a 50 m radius at 1 m resolution. A four-layer CNN produces 256 features, while a two-layer MLP encodes the goal vector into 128 features. Separate actor and critic heads each use 256 hidden units, with continuous actions bounded by $a_{\max} = 2$ m per coordinate. PPO (Schulman et al., 2017) uses discount $\gamma = 0.99$, GAE parameter 0.95, clipping threshold 0.2, four optimization epochs, four minibatches, and entropy coefficient 0.02. The value-loss coefficient and maximum gradient norm are both 0.5. The learning rate decays linearly from 2.5×10^{-4}. Each update collects 512 steps from each of eight parallel environments.

Navigation reward. Let d_t denote distance to the nearest point on the target building and b_{t+1} the nearest-obstacle distance after the action. Progress reward is $5(d_t - d_{t+1})$ and clearance reward is $-2\max(0, d_{\text{safe}} - b_{t+1})$, with $d_{\text{safe}} = 5$ m. Collision within $d_{\text{col}} = 2$ m gives -50, arrival within 5 m of the target building gives $+200$, and each step incurs -1. Distance milestones at 400, 300, 200, and 100 m each provide a one-time $+50$ reward.

Metric computation. GC and HS use a 2 m grid, with $\eta = 0.7$ for HS. Traversable cells account for each run's virtual obstacles. PDD uses 64-point arc-length resampling and all pairs from a random sample of at most 100 trajectories, normalized by resampled path length. EDE detects turns exceeding 90° after 64-point resampling and 21-point smoothing, then matches them to baseline junctions within 20 m and the closest exit direction. Unassigned trajectory sets use the baseline junction statistic. TPD uses five-point smoothing and heading-change thresholds of $\pm 15^\circ$.

Junction diagnostics and barriers. Diagnostics use a 5 m grid. Radial probes span 5–30 m in 5 m increments with 5° angular steps; valid candidates have at least three exits. DBSCAN merges detections within 20 m, and selected junctions are separated by at least 30 m. Barriers are 5 m thick and leave a 7–8 m gap on the wider free side, with clearance checks preventing sealed passages.

Heat and corridor obstacles. Heatmaps and route clustering use 5 m cells. Heat candidates are ranked by visitation with at least 30 m spacing. The overlap threshold for route clustering is adaptively raised from 0.5 to obtain additional candidates. Clusters are ranked by trajectory count, with the dominant set covering 80% of up to 4,000 sampled diagnostic trajectories. Circular radii range from 2 to 9 m according to Equation 6, and obstacles are rasterized on a 1 m grid.

Baselines. MBIE-EB (Strehl & Littman, 2008; Bellemare et al., 2016) counts visits to 5 m position cells paired with five heading-change bins and uses $\beta = 5\tau$. Random matches the combined layout's obstacle shapes, sizes, and count. Placements exclude targeted layout cells, avoid overlap between synthetic obstacles, and permit at most 10% overlap with building cells. Centers are sampled within $x \in [-1300, -750]$ m for AirSim16 and $x \in [-500, 490]$ m for AirSim23, using the scene's y range.