RECON: REPLAYABILITY-ENHANCED CLONING FOR OFF-TRAJECTORY NAVIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

A digital copy can look convincing and still give a robot the wrong map. This matters when a drone learns in a 3D clone built from a short scouting flight, then flies beyond the views used to build it. We study *task replayability*, the geometric reliability of a clone along later control routes. RECON represents this reliability with a Replayability Field $E : \mathcal{P} \to [0, 1]$ computed from refined Gaussians and scout poses with no learned parameters. On 3000 test-world poses, E has -0.58 Spearman correlation with depth error and detects high-gap geometry with 0.83 AUROC. The policy reads Gaussian tokens and E, while a support cost shapes its training routes. With architecture, source worlds, training budget, optimizer and seeds fixed, these components raise closed-loop success from 0.44 to 0.52 and reduce collisions from 0.56 to 0.48 on 600 episodes across 6 unseen reconstructed worlds. Across 40 real-UAV flights in 4 environments, RECON reaches 0.70 success, compared with 0.65 for the strongest image baseline.

1 INTRODUCTION

Real-to-sim pipelines reconstruct a scene from video, train a controller in the reconstruction and deploy it in the physical world. Novel-view synthesis reproduces source photographs (Mildenhall et al., 2021; Kerbl et al., 2023), feed-forward geometry recovers cameras, depth and points (Wang et al., 2024; 2025; Lin et al., 2025), and visuomotor systems use the resulting Gaussian scene as a training environment (Zhang et al., 2026; Torne et al., 2024; Qureshi et al., 2025). Control introduces a new requirement: the clone must preserve useful geometry along routes beyond the scout views.

A scout video $V_s = \{(I_t, p_t)\}_{t\in\mathcal{T}_s}$ produces a world $\mathcal{W} = f(V_s)$ by aligning renders to source frames,

$$\mathcal{L}_{\text{recon}}(\mathcal{W}) = \sum_{t\in\mathcal{T}_s} d\big(R(\mathcal{W}, p_t),\, I_t\big), \tag{1}$$

then a controller flies a route $\mathcal{T}_q$ that may approach from another side, start farther away or reverse the orbit. PSNR and LPIPS (Zhang et al., 2018) evaluate a render against an available photograph. Control depends on geometry where no reference photograph exists. We call the clone's validity along those later routes **task replayability**.

RECON (Replayability-Enhanced Cloning for Off-Trajectory Navigation) makes scout support explicit through one measurable chain (Fig. 1):

$$V_s \xrightarrow{(1)} \mathcal{W} \xrightarrow{(4),(5)} \{e_i\} \xrightarrow{(6)} E(p) \xrightarrow{(7),(8)} \pi_\theta, \qquad E(p) \text{ predicts } D_{\text{world}}(p) \text{ of Eq. (3).} \tag{2}$$

Each Gaussian stores the count and angular span of its scout observations. Their rendered support defines $E(p)$ at any reachable pose. The Gaussian Token Module (GTM) reads the clone, the Replayability Field Module (RFM) supplies e_i and $E(p)$ to the action model, and the Support Cost Module (SCM) favours training routes supported by the scout.

We contribute a protocol that measures replayability beyond scout views, a parameter-free field validated against depth and free-space errors, and a controlled study of its effect on navigation in simulation and on a real UAV. The four evaluations compare photometric quality with task replayability (RQ1), validate E against deployment-time proxies (RQ2), measure policy gains on unseen reconstructed worlds (RQ3), and test physical transfer (RQ4).

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.

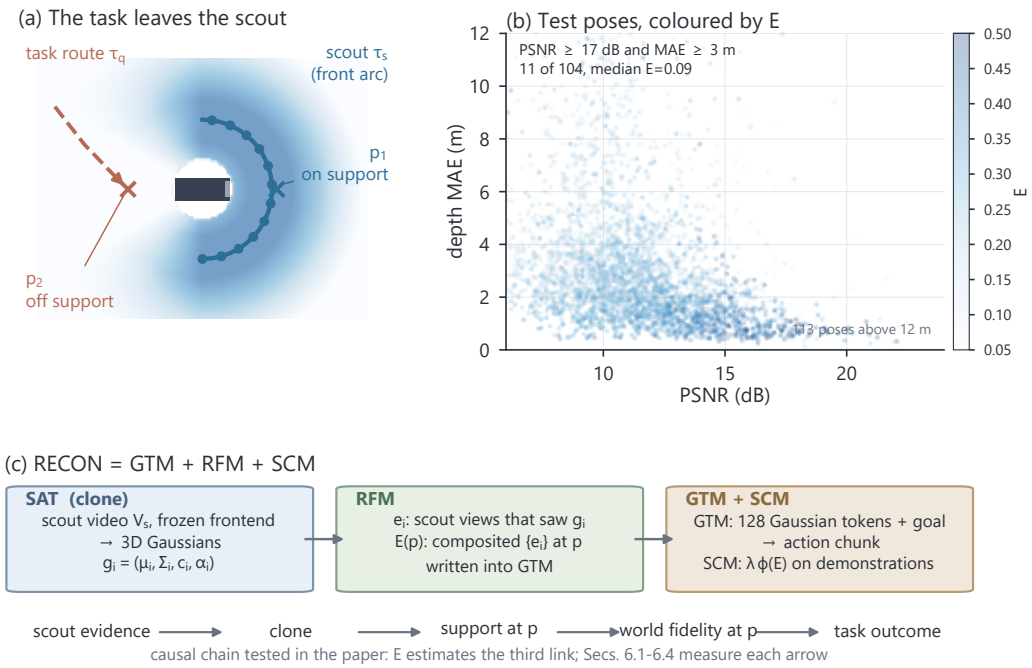

Figure 1: **Problem and method.** (a) A scout arc builds the clone and the task route leaves it. Query poses in the same clone receive different support values. (b) Test-world PSNR against depth MAE, coloured by E; the corner count marks poses with PSNR ≥ 17 dB and depth MAE ≥ 3 m. (c) GTM reads the clone, RFM renders support, and SCM assigns cost below the scout support.

2 Related Work

Photometric reconstruction and feed-forward geometry. Neural radiance fields and 3D Gaussian splatting synthesize views beyond the source cameras (Mildenhall et al., 2021; Kerbl et al., 2023). Later work improves aliasing, surfaces, anchors, initialization and motion (Yu et al., 2024; Huang et al., 2024a; Lu et al., 2024; Fan et al., 2024; Wu et al., 2024), with PSNR, SSIM and LPIPS as standard scores (Zhang et al., 2018). Feed-forward geometry recovers cameras and points from pairs (Wang et al., 2024; Leroy et al., 2024), sequences (Wang & Agapito, 2025) and many views (Yang et al., 2025; Wang et al., 2025; Yang et al., 2024a;b; Lin et al., 2025). RECON adds a pose-wise control validity measure to this reconstruction pipeline.

Uncertainty and pessimism. Bayes' Rays, FisherRF, density-aware ensembles and next-best-view planners estimate radiance-field uncertainty and validate it on held-out rendering error (Goli et al., 2024; Jiang et al., 2023; Sünderhauf et al., 2022; Jin et al., 2023). Model-based offline RL penalizes estimated model error during policy optimization (Yu et al., 2020; Kidambi et al., 2020). SCM applies the same principle to scout support in a real-to-sim clone.

Gaussian worlds as training and planning environments. World models generate rollouts in learned dynamics (Ha & Schmidhuber, 2018; Hafner et al., 2019; 2025). GaussFly trains UAV control in a Gaussian field (Zhang et al., 2026); RialTo and SplatSim train manipulation in reconstructed scenes (Torne et al., 2024; Qureshi et al., 2025); Gaussian SLAM builds these primitives online (Keetha et al., 2024; Matsuki et al., 2024; Yan et al., 2024; Huang et al., 2024b); and SplatNav and GaussNav plan on Gaussian maps (Chen et al., 2025; Lei et al., 2025). RECON augments the map with its scout support. Its controller uses flow matching (Lipman et al., 2022; Black et al., 2024), alongside comparisons with image policies and PPO (Schulman et al., 2017). AirSim supplies closed-loop depth and collision ground truth (Shah et al., 2017). These works provide the endpoints $V_s \to \mathcal{W}$ and $\mathcal{W} \to \pi_\theta$; the next section defines the geometric link between them.

3 TASK REPLAYABILITY

A scout records V_s around a stationary outdoor target, and the clone $\mathcal{W}$ minimizes Eq. (1) on its poses. A later route $\mathcal{T}_q$ ends at a standoff point 4.0 m from the target shell at 3.0 m height, in front of the target or behind it for reverse approaches. A query pose is any $p \in \mathcal{P}$ outside the scout trajectory. World-level replayability measures geometric agreement at p:

$$D_{\text{world}}(p) = d\big(G_{\mathcal{W}}(p),\, G_{\text{env}}(p)\big), \tag{3}$$

where G is the depth image and d is depth error or free-space disagreement. Task-level replayability is successful completion of $\mathcal{T}_q$ by a policy trained in $\mathcal{W}$. Thus Eq. (2) separates the link $\mathcal{W} \to D_{\text{world}}(p)$ from $D_{\text{world}}(p) \to \pi_\theta$; the method estimates the first with $E(p)$ and exposes it to the policy for the second.

4 METHOD

RECON factors Eq. (2) into a Support-Aware Twin (SAT), which maps the scout video to $(\mathcal{W}, E)$, and a Replay-Aware Action Model (RAM), which controls the vehicle in that twin. GTM forms Gaussian tokens, RFM computes per-Gaussian and pose support, and SCM adds support to the demonstration objective. Appendix B lists the fixed configuration.

4.1 SUPPORT-AWARE TWIN

Scout clone. AirSim provides poses and metric depth D_t; in the field, a frozen VGGT frontend (Wang et al., 2025) supplies poses, an initial point cloud and per-frame depth. The VGGT reconstruction is aligned to the flight-controller/GNSS trajectory through a Sim(3) transform, fixing scale and coordinate frame before computing E. A 3D Gaussian scene (Kerbl et al., 2023) with $g_i = (\mu_i, \Sigma_i, c_i, \alpha_i)$ and $c_i \in \mathbb{R}^3$ is refined for 7k steps at 960×540 under Eq. (1).

Per-Gaussian support. After refinement RFM writes a scalar e_i on each Gaussian from the scout poses. Gaussian i is *seen* by frame t if its centre projects inside the image at depth $z_{i,t} \in (0.2, 80)$ m and agrees with the frame's depth at that pixel,

$$s_{i,t} = \mathbf{1}\big[\, |z_{i,t} - D_t(u_{i,t})| < \max(0.45\,\text{m},\, 0.08\, D_t(u_{i,t}))\,\big], \tag{4}$$

which makes the count occlusion-aware. With $\mathbf{d}_{i,t}$ the unit direction from μ_i to camera t and $\bar{\mathbf{d}}_i$ the normalized mean of the seen directions,

$$n_i = \sum_{t \in \mathcal{T}_s} s_{i,t}, \qquad \theta_i = 2 \max_{t:\, s_{i,t}=1} \arccos\big(\mathbf{d}_{i,t} \cdot \bar{\mathbf{d}}_i\big) \quad (\theta_i = 0 \text{ if } n_i < 2), \tag{5}$$

$$e_i = \frac{\frac{1}{2}\, n_i/|\mathcal{T}_s| + \frac{1}{2}\min(\theta_i/\pi, 1)}{\max_j \big(\frac{1}{2}\, n_j/|\mathcal{T}_s| + \frac{1}{2}\min(\theta_j/\pi, 1)\big)} \in [0, 1].$$

n_i counts the scout views that constrained the primitive and θ_i is the angular span over which they did so; the two terms have equal weight, the normalization is by the best-supported Gaussian of the clone, and no parameter is learned. SAT outputs $(\mathcal{W}, \{e_i\})$.

Replayability Field Module. Rasterizing the Gaussians at p with e_i as the colour channel gives the premultiplied support $\hat{e}(u) = \sum_i T_i(u)\, \alpha_i(u)\, e_i$ and accumulated opacity $A(u) = \sum_i T_i(u)\, \alpha_i(u)$ at pixel u, with T_i the front-to-back transmittance. The field averages support over rendered opacity,

$$\boxed{E(p) = \frac{\sum_u \hat{e}(u)}{\sum_u A(u)} \in [0, 1]}, \tag{6}$$

with $E(p) = 0$ when $\sum_u A(u) = 0$. Empty pixels contribute zero, and covered pixels contribute in proportion to opacity. One rasterization from the task camera gives $E(p)$ for calibration and control.

Threshold and low-support cost. The threshold τ is the 20th percentile of $\{E(p_t) : t \in \mathcal{T}_s\}$ for the same clone. The high-support set $\mathcal{E}_\tau = \{p : E(p) \geq \tau\}$ therefore contains at least 80% of its scout poses. The cost is zero on $\mathcal{E}_\tau$ and linear below it,

$$\phi(E) = \max(0,\, \tau - E). \tag{7}$$

4.2 REPLAY-AWARE ACTION MODEL

Observation. At pose p with position x_p, let $s_i = (\det \Sigma_i)^{1/6}$ be the Gaussian's scalar equivalent scale. GTM reads the 128 nearest Gaussians as $f_i = [\mu_i - x_p, \log s_i, c_i, \alpha_i, e_i] \in \mathbb{R}^9$. The conditioning vector $h = [\xi, g - x_p, E(p), a^{(\xi)}]$ contains flow time, relative goal, rendered support and the noisy action chunk. Closing RFM sets e_i and $E(p)$ to zero. The pose comes from AirSim or from the flight controller state registered to the scout track.

Architecture and training. An 8-layer pre-norm transformer encoder (width 576, 8 heads, feed-forward 2304) reads the projected tokens with h added to each. A two-layer head on the mean-pooled output predicts 8 three-dimensional displacements, and the controller executes the first 4 before replanning. Conditional flow matching (Lipman et al., 2022) uses $a^{(\xi)} = (1 - \xi)\epsilon + \xi a$ with target $a - \epsilon$; inference uses 4 Euler steps. The model has 32.3M trainable parameters.

Support Cost Module. RAM distils a planner that runs inside the clone, and SCM is the term $\lambda\,\phi(E)$ in that planner. From a start pose the planner searches a $0.5\,\mathrm{m}$ lattice for the path to the standoff point that minimizes

$$J(\mathcal{T}_q) = \sum_k \Big[\|a_k\| + \lambda\,\phi\big(E(p_k)\big)\Big] \quad \text{s.t. clearance} \geq 4.0\ \mathrm{m} \text{ from the target shell in } \mathcal{W}, \tag{8}$$

and GTM is trained to reproduce the resulting trajectories: shortest clearance-respecting paths with SCM closed ($\lambda = 0$), paths that trade length for time above the scout's support with SCM open ($\lambda = 0.4$). λ and the percentile of τ are chosen on the validation worlds.

The three rows. The causal comparison matches the training worlds and demonstration starts, architecture, parameter count, optimizer, training steps and seeds while opening the modules in order; SCM changes the routes demonstrated from those starts. **+GTM** zeroes e_i and $E(p)$ and sets $\lambda = 0$. **+GTM+RFM** supplies e_i and $E(p)$ with $\lambda = 0$. **+GTM+RFM+SCM (RECON)** sets $\lambda = 0.4$ in Eq. (8). The composed system is

$$(\mathcal{W}, E) = \mathrm{SAT}(V_s), \qquad \pi_{\mathrm{replay}}(p, g) = \pi_\theta(\mathrm{GTM}(\mathcal{W}, p), E(p), g)\,, \tag{9}$$

where SCM determines the demonstrations used to fit π_θ. The experiments now test the two remaining links, $E(p) \leftrightarrow D_{\mathrm{world}}(p)$ and $\pi_{\mathrm{replay}} \rightarrow \mathrm{Success}(\mathcal{T}_q)$.

5 EXPERIMENTAL DESIGN

Reconstructed worlds. A *world* is one reconstructed scene and target with its own scout video and clone. We use 16 AirSim worlds (Shah et al., 2017) from 4 layout families, split into 8 training, 2 validation and 6 test worlds (Fig. 2b, Appendix C). Training worlds supply demonstrations, validation worlds select τ, λ and checkpoints, and test worlds supply the headline results.

Scout pattern and route regimes. In a target-centred frame, r is horizontal distance to the target shell and ψ is azimuth from its front axis. The scout is the front half of a standoff orbit, $|\psi| \leq 90^\circ$ and $r = 4.0\,\mathrm{m}$, flown at 3.0 and $4.5\,\mathrm{m}$ with the camera on the target. Rear frames are held out for measurement. Four route regimes vary the start sector and stop point (Fig. 2a; Appendix D): Near starts inside the scout arc; Moderate cuts in laterally; Large passes a weakly observed flank; Reverse starts and stops behind the target. Along their reference paths, mean E falls from Near to Reverse and the fraction below τ rises from 0.06 to 0.75.

Each world contributes 25 fixed starts per regime, giving 600 test episodes per method. An episode ends at success, contact or 70 steps of at most $1.2\,\mathrm{m}$. Success requires stopping within $1.5\,\mathrm{m}$ horizontally of the standoff point, inside the 2–4 m height band and without contact. Collision denotes target contact or a simulator collision flag. Results are averaged over test worlds with scene-level bootstrap statistics; policies train in the clone and execute in AirSim.

Query-pose calibration set. Each world supplies 500 query poses over $r \in [4, 14]\,\mathrm{m}$, all azimuths, heights 2–6 m and $\pm 30^\circ$ yaw jitter: 8000 poses in total and 3000 on test worlds (Appendix E). Each

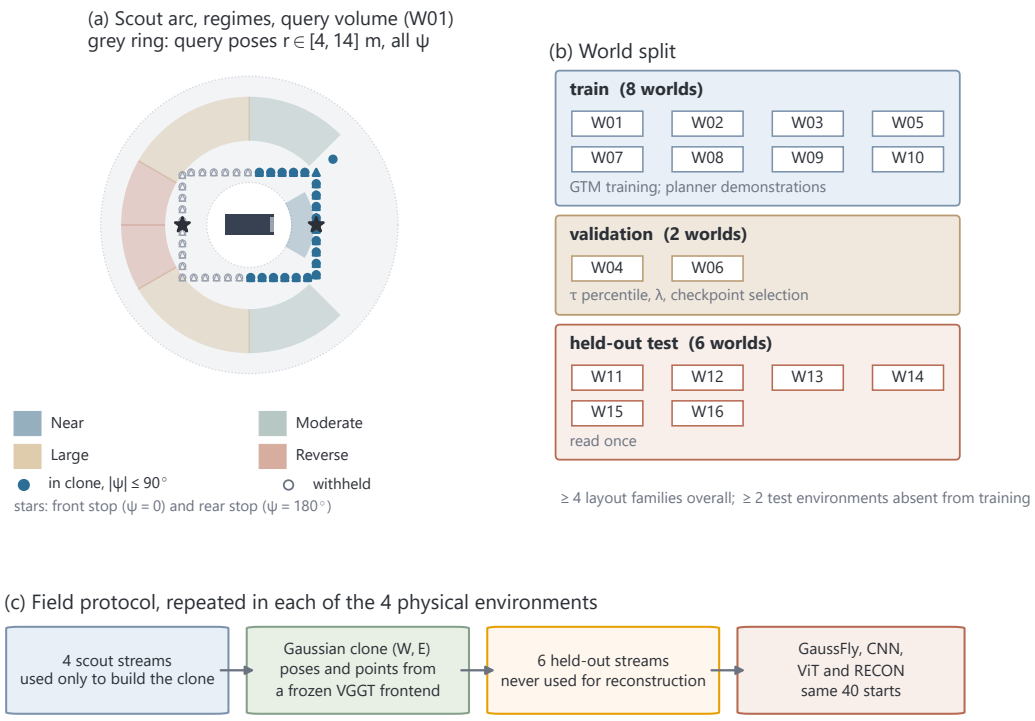

Figure 2: **Experimental construction.** (a) Scout poses of AirSimNH–Sedan–01: filled markers build the clone; hollow markers are held out. Shading marks four start regimes, stars mark standoff points, and the grey annulus is the query volume. (b) World split. (c) Field protocol in each of 4 environments.

pose stores $E(p)$; nearest-scout distance $d_s(p)$, supporting view count $n_v(p)$ and angular coverage $\Delta\psi(p)$; PSNR and LPIPS; depth MAE; free-space disagreement at 25 m; and a high-gap flag for depth MAE above 2.0 m. We report Spearman correlation, high-gap AUROC/AUPRC and route-level AUROC for +GTM failure, with 1000 bootstrap samples stratified by world.

Policies. Table 2 compares six controllers on the identical start set: a GaussFly-style *CNN+PPO* (Zhang et al., 2026; Schulman et al., 2017) on 224×224 clone renders, trained for at least 1.0M interactions; two *2D flow-matching* heads on the same renders with the same flow objective, demonstrations and seeds as RAM, a CNN (0.46M) and a ViT matched to GTM in depth and width (32.50M); and the three RAM rows. All flow-matching heads see 40k planner trajectories from the training worlds and train for 200k AdamW steps on 3 seeds (Appendix B).

Field protocol. The physical study uses 4 environments, each with an independent scout, clone and held-out routes (Appendix F). Sample collection and algorithm deployment both run on the platform in Fig. 3. Aerial orbits at 3 and 4.5 m together with ground views build $(\mathcal{W}, E)$ through frozen VGGT. Held-out routes are registered to the scout frame where registration succeeds, and per-frame E and LPIPS use those frames. GaussFly, the 2D CNN, the 2D ViT and RECON fly the same 40 starts. Table 3 reports success, collision or intervention, trajectory cost, latency and mean E. Following Eq. (2), simulation query poses test $E(p) \leftrightarrow D_{\text{world}}(p)$ in RQ1–RQ2, and the simulation and field routes test $\pi_{\text{replay}} \rightarrow \text{Success}(\mathcal{T}_q)$ in RQ3–RQ4.

6 Results

6.1 Phenomenon (RQ1)

Figure 4 follows Eq. (2) from image quality through geometry to control. From Near to Reverse, PSNR drops by 3.68 dB while depth MAE grows from 1.52 m to 4.23 m. Across test-world regimes,

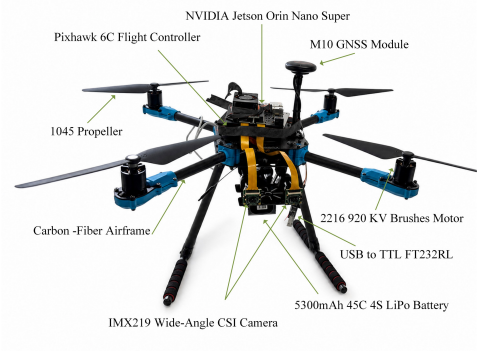

Figure 3: **Field platform.** Scout collection and policy deployment both use this carbon-fiber multi-rotor (Pixhawk 6C, M10 GNSS, IMX219 wide-angle camera, Jetson Orin Nano Super).

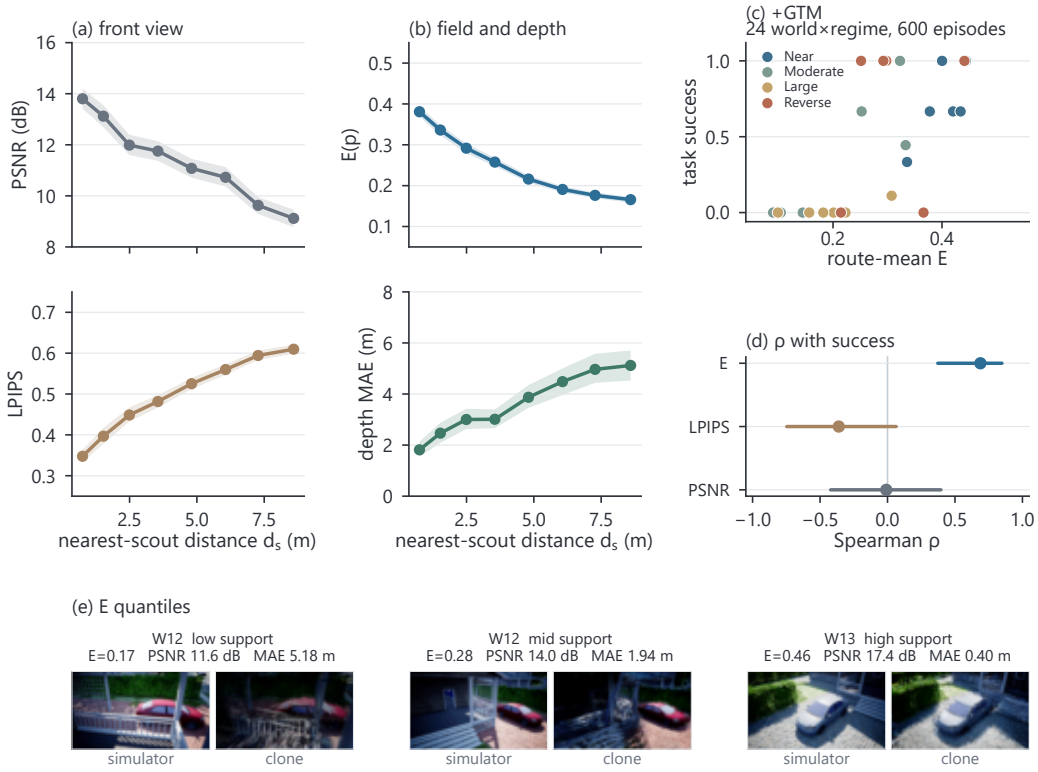

Figure 4: **Phenomenon (RQ1).** (a) PSNR and LPIPS and (b) E and depth MAE against nearest-scout distance on front-view test-world query poses ($|\psi| \leq 90°$); equal-count bins, bootstrap 95% bands. (c) Closed-loop success of +GTM against route-mean E, one point per test world and regime (600 episodes). (d) Spearman correlation of that success with PSNR, LPIPS and route-mean E. (e) Simulator photograph and clone render at three E quantiles with measured E, PSNR and depth MAE.

+GTM success has Spearman correlation -0.01 with PSNR, -0.36 with LPIPS and 0.69 with route-mean E (panels c,d). Panel (e) shows the corresponding visual and geometric changes from the same camera.

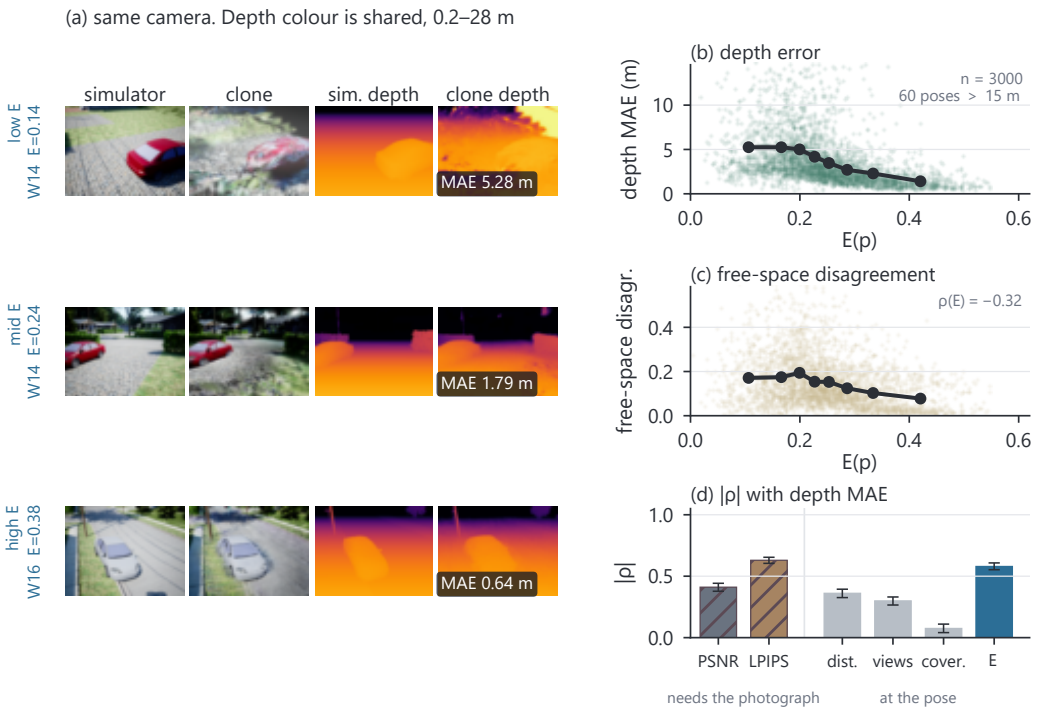

Figure 5: **Validation of the field (RQ2).** (a) Simulator and clone RGB and depth at three support levels; depth uses a shared 0.2–28 m scale. (b,c) Depth MAE and free-space disagreement against E on 3000 test poses, with equal-count bins and bootstrap 95% bands. (d) Absolute depth-MAE correlation; E is strongest among scores available without a reference photograph.

Table 1: **Replayability validity (RQ2)** on 3000 test poses. We report Spearman ρ, high-gap AUROC/AUPRC and route-failure AUROC. Route failure uses minimum E and the mean of each other predictor. Gray $\pm$ is bootstrap SD.

Predictor	ρ depth MAE	ρ free space	$\text{AUROC}_{\text{gap}}$	$\text{AUPRC}_{\text{gap}}$	$\text{AUROC}_{\text{fail}}$
PSNR	-0.41 $\pm$0.02	-0.51 $\pm$0.02	0.73 $\pm$0.01	0.82 $\pm$0.01	0.49 $\pm$0.03
LPIPS (Zhang et al., 2018)	0.63 $\pm$0.01	0.46 $\pm$0.02	0.82 $\pm$0.01	0.87 $\pm$0.01	0.68 $\pm$0.04
Nearest-scout distance	0.36 $\pm$0.02	0.32 $\pm$0.02	0.70 $\pm$0.01	0.77 $\pm$0.01	0.51 $\pm$0.03
Supporting view count	-0.30 $\pm$0.02	-0.33 $\pm$0.02	0.65 $\pm$0.01	0.74 $\pm$0.01	0.58 $\pm$0.04
Angular coverage	0.07 $\pm$0.02	0.07 $\pm$0.02	0.53 $\pm$0.01	0.67 $\pm$0.01	0.53 $\pm$0.04
Replayability field E	-0.58 $\pm$0.01	-0.32 $\pm$0.02	0.83 $\pm$0.01	0.87 $\pm$0.01	0.87 $\pm$0.02

6.2 Validation of the Field (RQ2)

Figure 5 calibrates E on 3000 test-world poses, and Table 1 compares all predictors. The Spearman correlation of E with depth MAE is -0.58 $\pm$0.01 and with free-space disagreement is -0.32 $\pm$0.02. Nearest-scout distance reaches 0.36 on depth MAE, while reference-based LPIPS reaches 0.63. High-gap AUROC is 0.83 for E and 0.70 for nearest-scout distance. Route-minimum E predicts +GTM failure with 0.87 AUROC. Among deployment-time scores, E leads four of five criteria and is within 0.01 of the best free-space correlation.

6.3 Exploitation by the Policy (RQ3)

Table 2 compares 600 episodes per method across 6 worlds and 3 seeds. In the causal ablation, opening RFM and SCM raises overall success from 0.44 to 0.52 and lowers collisions from 0.56 to 0.48. Near success rises from 0.72 to 0.89. Across the three off-support regimes, success rises from

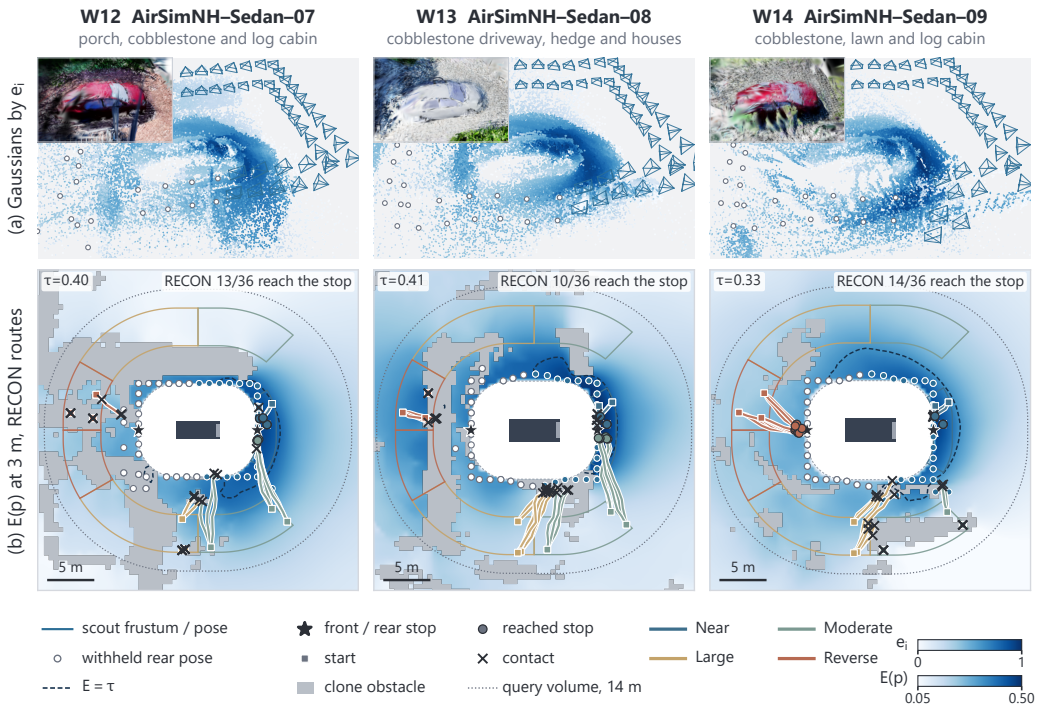

Figure 6: **Scout support and RECON routes in three test worlds.** (a) Gaussian centres coloured by e_i, seen from above the flank with the front to the right; pyramids are the scout cameras that built the clone, hollow circles are the withheld rear poses, and the inset renders the same view in RGB, zoomed on the target. (b) $E(p)$ at 3 m on the policy lattice, in the colour map of Fig. 1b. Dashed: $E = \tau$; grey: clone obstacles; white: 4 m shell clearance; outlines: start sectors; stars: stops. Lines are RECON rollouts in AirSim from the first three starts of each regime.

Table 2: **Main closed-loop comparison (RQ3).** Success and collision on 600 episodes per method, averaged over 3 seeds. Overall is the unweighted regime mean; gray ± is scene-level bootstrap SD. The last rows open GTM, RFM and SCM in order. Parameter counts are rounded in the table; the 2D CNN has 0.46M parameters.

		Task success ↑					
Method	Params	Near	Mod.	Large	Rev.	Overall ↑	Collision ↓
GaussFly (Zhang et al., 2026)	0.16M	0.64±0.06	0.18±0.04	0.08±0.00	0.50±0.03	0.35±0.04	0.62±0.05
2D flow matching (Lipman et al., 2022)							
CNN	0.5M	0.80±0.04	0.19±0.08	0.00±0.00	0.61±0.18	0.40±0.03	0.60±0.03
ViT	32.5M	0.83±0.07	0.33±0.14	0.08±0.06	0.61±0.16	0.46±0.07	0.54±0.06
+GTM	32.3M	0.72±0.10	0.35±0.16	0.02±0.01	0.65±0.17	0.44±0.05	0.56±0.08
+GTM+RFM	32.3M	0.81±0.09	0.37±0.16	0.06±0.04	0.67±0.17	0.48±0.06	0.51±0.06
+GTM+RFM+SCM (RECON)	32.3M	0.89±0.06	0.39±0.15	0.11±0.09	0.67±0.17	0.52±0.08	0.48±0.05

0.34 to 0.39. Every regime improves or holds at each step. The matched ViT and +GTM differ by 0.02 overall, so this comparison isolates the representation; the step from +GTM to +GTM+RFM isolates the replayability features.

Figure 6 shows the support field of three test worlds with RECON rollouts re-flown in AirSim. Per-Gaussian support is highest on the scout-facing side of each clone, and $E(p)$ at flight height falls from the front arc toward the flank and the rear.



Figure 7: **Replayability on recorded viewpoints.** Each row is one site; photograph and RECON reconstruction share a camera (scout left pair, held-out right pair). Corner value: measured E.

Table 3: **Real UAV (RQ4).** Forty shared starts across 4 environments. Failure combines collision and pilot intervention; cost is path length divided by planned length. Latency is measured on an RTX 4080 SUPER. The image policies render at 224^2; RECON gathers 128 tokens and renders E at 256×144.

Method	Success ↑	Failure ↓	Cost ↓	ms/frame ↓	Mean E
GaussFly (Zhang et al., 2026)	0.425	0.35	1.09	2.7	0.21
2D flow matching (Lipman et al., 2022)					
CNN	0.55	0.20	1.13	3.7	0.28
ViT	0.65	0.20	1.06	21.0	0.33
RECON	0.70	0.15	1.15	12.4	0.37

6.4 Transfer to a Real UAV (RQ4)

Figure 7 and Table 3 report 4 physical environments flown on the platform of Fig. 3. Route-mean E falls from 0.40 on near routes to 0.16 on reverse routes, and its per-frame correlation with held-out-frame LPIPS is -0.45. Every method flies the same 40 starts. RECON reaches 0.70 success, compared with 0.65 for the strongest image baseline, and 0.15 collisions or interventions. Appendix F shows recorded frames.

Figure 9 in Appendix F places the recorded tracks on the field $E(p)$. The F03 along-body track stays above τ, while 44% of the shown F01 held-out frames and 24% of the F04 held-out frames fall below it; by construction of τ, 20% of scout frames do.

7 Conclusion

Visual similarity alone does not locate reliable control geometry. RECON maps scout evidence into a Replayability Field, validates the field against world geometry and exposes it to the policy. Simulation and real-UAV results complete the chain $V_s \to \mathcal{W} \to E(p) \to \pi_\theta$ from Eq. (2). Photometric fidelity evaluates rendering, while task replayability evaluates the clone along control routes.

Reproducibility statement

Separate directories preserve the scout and held-out splits. SAT refines one Gaussian world per scene and target; RFM follows Eqs. (4)–(6); and SCM follows Eq. (8). Appendix B gives all hyperparameters. The release contains fixed world splits, starts, pose sampling, evaluation bins, field catalogues, recording scripts and flight logs.

Ethics statement

The work uses a simulator and outdoor flights around parked, unoccupied vehicles on closed roads with a safety pilot holding the controls. Every frame shown in the paper is free of people. A clone trusted outside its scout support can produce unsafe flight, and the Replayability Field makes that trust visible.

AI use statement

During the preparation of this work, the author(s) used Overleaf's built-in AI grammar checking tool, Writefull for Overleaf, to identify and correct grammatical errors and improve readability. The author(s) used Grok 4.7 to assist with the experiments. After using these tools, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

References

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Timothy Chen, Ola Shorinwa, Joseph Bruno, Aiden Swann, Javier Yu, Weijia Zeng, Keiko Nagami, Philip Dames, and Mac Schwager. Splat-nav: Safe real-time robot navigation in gaussian splatting maps. *IEEE Transactions on Robotics*, 41:2765–2784, 2025.

Zhiwen Fan, Wenyan Cong, Kairun Wen, Kevin Wang, Jian Zhang, Xinghao Ding, Danfei Xu, Boris Ivanovic, Marco Pavone, Georgios Pavlakos, et al. Instantsplat: Sparse-view gaussian splatting in seconds. *arXiv preprint arXiv:2403.20309*, 2024.

Lily Goli, Cody Reading, Silvia Sellán, Alec Jacobson, and Andrea Tagliasacchi. Bayes' rays: Uncertainty quantification for neural radiance fields. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20061–20070. IEEE, 2024.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *Advances in neural information processing systems*, 31, 2018.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025.

Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 conference papers*, pp. 1–11, 2024a.

Huajian Huang, Longwei Li, Hui Cheng, and Sai-Kit Yeung. Photo-slam: Real-time simultaneous localization and photorealistic mapping for monocular, stereo, and rgb-d cameras. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21584–21593. IEEE, 2024b.

Wen Jiang, Boshu Lei, and Kostas Daniilidis. Fisherrf: Active view selection and uncertainty quantification for radiance fields using fisher information. *arXiv preprint arXiv:2311.17874*, 2023.

Liren Jin, Xieyuanli Chen, Julius Rückin, and Marija Popović. Neu-nbv: Next best view planning using uncertainty estimation in image-based neural rendering. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 11305–11312. IEEE, 2023.

Nikhil Keetha, Jay Karhade, Krishna Murthy Jatavallabhula, Gengshan Yang, Sebastian Scherer, Deva Ramanan, and Jonathon Luiten. Splatam: Splat, track & map 3d gaussians for dense rgb-d slam. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21357–21366. IEEE, 2024.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis, et al. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. Morel: Model-based offline reinforcement learning. *Advances in neural information processing systems*, 33: 21810–21823, 2020.

Xiaohan Lei, Min Wang, Wengang Zhou, and Houqiang Li. Gaussnav: Gaussian splatting for visual navigation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):4108–4121, 2025.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European conference on computer vision*, pp. 71–91. Springer, 2024.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Tao Lu, Mulin Yu, Linning Xu, Yuanbo Xiangli, Limin Wang, Dahua Lin, and Bo Dai. Scaffold-gs: Structured 3d gaussians for view-adaptive rendering. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20654–20664. IEEE, 2024.

Hidenobu Matsuki, Riku Murai, Paul HJ Kelly, and Andrew J Davison. Gaussian splatting slam. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18039–18048. IEEE, 2024.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

M Nomaan Qureshi, Sparsh Garg, Francisco Yandun, David Held, George Kantor, and Abhisesh Silwal. Splatsim: Zero-shot sim2real transfer of rgb manipulation policies using gaussian splatting. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6502–6509. IEEE, 2025.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and service robotics: Results of the 11th international conference*, pp. 621–635. Springer, 2017.

Niko Sünderhauf, Jad Abou-Chakra, and Dimity Miller. Density-aware nerf ensembles: Quantifying predictive uncertainty in neural radiance fields. *arXiv preprint arXiv:2209.08718*, 2022.

Marcel Torne, Anthony Simeonov, Zechu Li, April Chan, Tao Chen, Abhishek Gupta, and Pulkit Agrawal. Reconciling reality through simulation: A real-to-sim-to-real approach for robust manipulation. *arXiv preprint arXiv:2403.03949*, 2024.

Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. In *2025 International Conference on 3D Vision (3DV)*, pp. 78–89. IEEE, 2025.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306. IEEE, 2025.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709. IEEE, 2024.

Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20310–20320. IEEE, 2024.

Chi Yan, Delin Qu, Dan Xu, Bin Zhao, Zhigang Wang, Dong Wang, and Xuelong Li. Gs-slam: Dense visual slam with 3d gaussian splatting. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19595–19604. IEEE, 2024.

Jianing Yang, Alexander Sax, Kevin J Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21924–21935. IEEE, 2025.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10371–10381. IEEE, 2024a.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in neural information processing systems*, 37:21875–21911, 2024b.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. *Advances in neural information processing systems*, 33:14129–14142, 2020.

Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19447–19456. IEEE, 2024.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Yuhang Zhang, Mingsheng Li, Yujing Shang, Zhuoyuan Yu, Chao Yan, Jiaping Xiao, and Mir Feroskhan. Gaussfly: Contrastive reinforcement learning for visuomotor policies in 3d gaussian fields. *arXiv preprint arXiv:2604.05062*, 2026.

A Algorithm

B Exact Configuration

Table 4 lists every fixed quantity of the protocol.

C World Catalogue

Table 5 lists every reconstructed world with its academic name, raw simulator identifier, target, layout and split. The scout column reports the number of scout frames that enter the clone (front arc, two heights) and the 20th-percentile scout E that becomes τ.

Algorithm 1 RECON

Require: Scout video $V_s = \{(I_t, p_t)\}_{t \in \mathcal{T}_s}$.

1: Refine $\mathcal{W}$ by minimizing $\mathcal{L}_{\text{recon}}$ of Eq. (1) on $p_t \in \mathcal{T}_s$.
2: RFM: for each Gaussian, count seen frames n_i and angular span θ_i from the scout poses (Eqs. (4)–(5)); write e_i.
3: RFM: define $E(p)$ as the opacity-normalized render of $\{e_i\}$ (Eq. (6)).
4: Set τ to the 20th percentile of E on $\mathcal{T}_s$; set $\phi(E) = \max(0, \tau - E)$ (Eq. (7)).
5: Generate planner demonstrations in $\mathcal{W}$ minimizing Eq. (8) with SCM open ($\lambda = 0.4$, RECON) or closed ($\lambda = 0$).
6: Train GTM with RFM channels $(e_i, E(p))$ (+GTM+RFM and RECON) or with those channels zeroed (+GTM).
7: Fly held-out routes in the environment; report success and collision per regime.

Table 4: Frozen configuration.

Clone (SAT) and RFM	
Frontend	frozen VGGT (field: poses, points, depth); simulator poses and depth (AirSim)
Gaussian refinement	7k steps at 960×540; densify and prune as in the released code
Support e_i	RFM, Eqs. (4)–(5): seen test $\lvert z - D\rvert < \max(0.45\,\text{m}, 0.08D)$, equal weights, max-normalized
Field $E(p)$	RFM, Eq. (6), task camera on the target centre, 256×144 raster
Threshold τ	20th percentile of E on the scout poses of the same clone
SCM	$\phi(E) = \max(0, \tau - E)$, $\lambda = 0.4$ when SCM is open (validation worlds)
GTM (RAM)	
Tokens	GTM: 128 nearest Gaussians, $s_i = (\det \Sigma_i)^{1/6}$, 9-d features $[\mu_i - x_p, \log s_i, c_i, \alpha_i, e_i]$
Conditioning	flow time, goal-relative vector, $E(p)$, noisy chunk
Encoder	8 pre-norm layers, width 576, 8 heads, FF 2304, GELU, mean pool
Action	chunk of 8 3-D displacements (≤ 1.2 m each), execute 4
Flow matching	linear interpolant, MSE on velocity, 4 Euler steps at inference
Parameters	32.3M (exact 32.28M)
Optimizer	AdamW, lr 3×10^{-4}, batch 256, weight decay 0.01, cosine decay, 200k steps
Data	40k planner trajectories, 300k windows, 8 training worlds
Seeds	3
Baselines	
2D flow matching (Lipman et al., 2022)	224^2 clone renders. CNN, 0.46M. ViT matched to GTM in depth and width, 32.50M. Same flow objective, data, steps and seeds
CNN+PPO (GaussFly-style) (Zhang et al., 2026; Schulman et al., 2017)	224^2 clone renders, 0.16M, ≥ 1.0M interactions, extended while improving
Reference path	lattice search of Eq. (8) on simulator ground-truth occupancy, $\lambda = 0$ (route support statistics only)
Evaluation	
Success	within 1.5 m of the standoff, height 2–4 m, no contact, ≤ 70 steps
Episodes	25 per regime per world; identical start set for every method
Query poses	500 per world, $r \in [4, 14]$ m, $h \in [2, 6]$ m, yaw jitter $\pm 30^\circ$
World gap	depth MAE, free-space disagreement at 25 m, high gap = MAE > 2.0 m
Statistics	Spearman, AUROC/AUPRC, 1000-sample bootstrap stratified by world

D Route Regimes and Construction

Table 6 gives the exact geometry of the four regimes and the support measured on their routes. Start poses are drawn per world from the regime sectors with a fixed seed: r uniform in the band, ψ uniform in the sector (mirrored across the front axis for Moderate and Large), height uniform in $[2.5, 3.5]$ m, camera on the target centre. Starts that fall inside simulator geometry or within 2.2 m of the shell are resampled. The reference path of every start is the lattice search of Eq. (8) with $\lambda = 0$

Table 5: Reconstructed worlds. Names follow *Environment–Target–Instance*; raw identifiers are kept for the released code.

ID	Name	Raw identifier	Target	Layout	Split	Scout frames	Gaussians	τ
W01	AirSimNH–Sedan–01	Car_107	dark sedan	open suburban street, lawn and trees	train	42	51k	0.22
W02	AirSimNH–Sedan–02	Car_42	silver sedan	cobblestone driveway under trees	train	42	58k	0.26
W03	AirSimNH–Sedan–03	Car_64	dark sedan	hedge-lined suburban street	train	42	52k	0.20
W04	AirSimNH–Sedan–04	Car_100	red sedan	hedge street, foliage-occluded orbit	val	42	58k	0.12
W05	CityEnviron–Sedan–01	Vehicle_A_182	white sedan	multi-lane urban avenue	train	42	55k	0.22
W06	CityEnviron–PoliceCruiser–01	Vehicle_Police7_235	police cruiser	urban intersection, street tree	val	42	59k	0.23
W07	CityEnviron–Taxi–01	Vehicle_Taxi32	yellow taxi	urban street, parked traffic	train	42	59k	0.27
W08	AirSimNH–Sedan–05	Car_14	light sedan	carport, picket fence and houses	train	42	66k	0.36
W09	AirSimNH–Sedan–06	Car_35	sedan	hedge, stone wall and houses	train	42	65k	0.37
W10	CityEnviron–Taxi–02	Vehicle_Taxi12	yellow taxi	multi-lane street, glass office, street trees	train	42	66k	0.41
W11	Coastline–Path–01	VehicleAdvExampleMap_C_0	path	curved path, boulder, palms and ferns	test	42	71k	0.48
W12	AirSimNH–Sedan–07	Car_132	sedan	porch, cobblestone and log cabin	test	42	64k	0.40
W13	AirSimNH–Sedan–08	Car_21	sedan	cobblestone driveway, hedge and houses	test	42	69k	0.41
W14	AirSimNH–Sedan–09	Car_49	sedan	cobblestone, lawn and log cabin	test	42	66k	0.33
W15	AirSimNH–Sedan–10	Car_56	sedan	spruce-occluded street, stone wall	test	42	71k	0.29
W16	AirSimNH–Sedan–11	Car_74	sedan	grass verge, sidewalk and porch house	test	42	66k	0.38

on ground-truth occupancy; its route-mean and route-minimum E, fraction below τ and maximum nearest-scout distance are the support statistics of Table 6.

Table 6: **Route regimes.** Construction columns are fixed; support columns are measured on the ground-truth reference path of every held-out episode (mean over test worlds, gray $\pm$ scene-level bootstrap SD). r in metres to the target shell, ψ the azimuth from the front axis; Large passes the flank at $r \geq 5.5\,\mathrm{m}$ before turning to the front stop. $d_{\max}$ is the largest nearest-scout distance on a route, averaged over routes and test worlds.

	Construction			Measured support of the route			
Regime	start r	start ψ	stop	mean E	min E	frac. $E < \tau$	$d_{\max}$ (m)
Near	$[4, 6]$	$\lvert\psi\rvert \leq 30^\circ$	front	0.46$\pm$0.01	0.43$\pm$0.01	0.06$\pm$0.03	1.1$\pm$0.0
Moderate	$[8, 12]$	$45^\circ \leq \lvert\psi\rvert \leq 90^\circ$	front	0.34$\pm$0.01	0.12$\pm$0.02	0.51$\pm$0.04	5.3$\pm$1.0
Large	$[8, 12]$	$90^\circ < \lvert\psi\rvert \leq 150^\circ$	front, via flank	0.34$\pm$0.02	0.15$\pm$0.03	0.55$\pm$0.06	8.7$\pm$0.7
Reverse	$[8, 12]$	$150^\circ < \lvert\psi\rvert \leq 180^\circ$	rear	0.31$\pm$0.03	0.26$\pm$0.02	0.75$\pm$0.15	12.5$\pm$0.1

E Query-Pose Calibration Set

For each world the sampler draws 500 poses: horizontal distance to the shell $r \sim U[4, 14]\,\mathrm{m}$, azimuth $\psi \sim U[-180^\circ, 180^\circ]$, height $h \sim U[2, 6]\,\mathrm{m}$, yaw toward the target centre plus $U[-30^\circ, 30^\circ]$, pitch from the height geometry to a look-at point 1 m above ground. Poses whose simulator depth at the image centre is below $0.5\,\mathrm{m}$ or that collide are resampled. The stored record per pose is: pose;

$E(p)$; $d_s(p)$; $n_v(p)$; $\Delta\psi(p)$; render PSNR and LPIPS against the simulator photograph; clone depth and simulator depth at 256×144; depth MAE; free-space disagreement; high-gap flag.

F FIELD ENVIRONMENTS

Table 7 catalogues the physical environments. F01–F04 were recorded on farmland roads with a red compact SUV, an empty flatbed truck, a dump truck and a white mid-size SUV as targets; they occupy different road segments with different surroundings. Each environment contributes four scout streams and six held-out streams. Figure 8 places scout, held-out and pass-by frames in three rows; the columns are the red SUV, the blue flatbed, the dump truck and the white SUV.

Table 7: Physical environments. Scout streams build the clone; held-out streams are evaluation only.

ID	Environment	Target	Layout	Scout data	Held-out data
F01	Farmland road, compact SUV	red short-wheelbase SUV	yellow-line concrete road, paddy and drainage channel	4	6
F02	Tree-lined road, flatbed truck	blue empty flatbed truck	concrete road between tree rows	4	6
F03	Pond-side road, dump truck	white-cab dump truck	concrete road, roadside pond and trees	4	6
F04	Open shoulder, white SUV	white mid-size SUV	paved shoulder next to dry ditch and farmland	4	6



Figure 8: **Recorded field frames.** Each row is one view type: scout, held-out, pass-by. Each column is one vehicle: red SUV, blue flatbed, dump truck, white SUV.

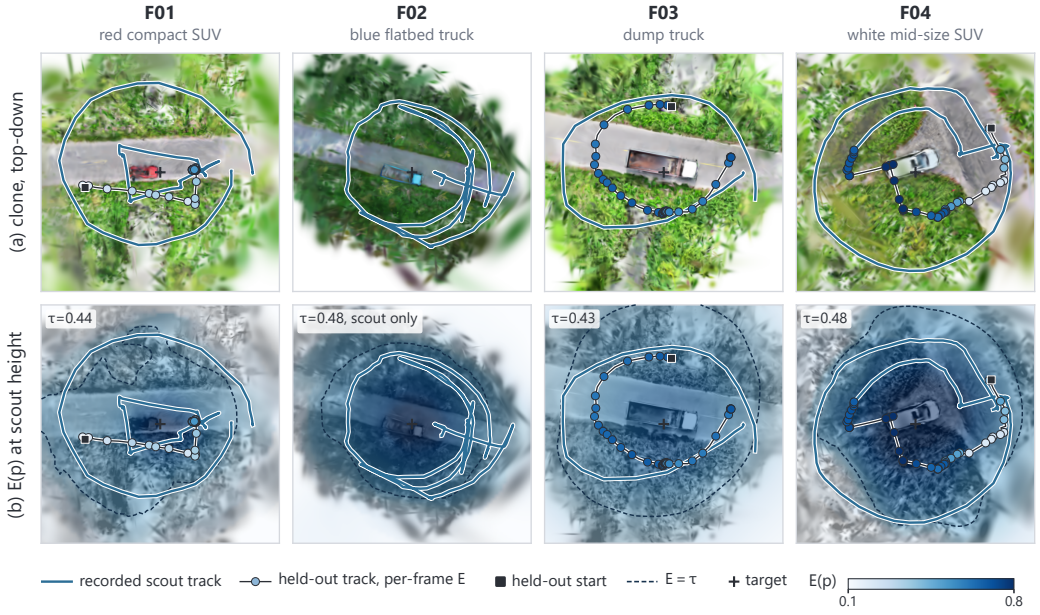

Figure 9: **Recorded field routes on the RECON field.** (a) Each field clone splatted from above. (b) $E(p)$ on a horizontal lattice at median scout height, camera on the target (colour scale 0.1–0.8), over the clone in grey; dashed: $E = \tau$. Blue: recorded scout track. Dots: registered held-out track coloured by per-frame E; the square is its first frame. F01 shows its longest continuously registered segment, the last 97 of 257 held-out frames; F02 has no registered held-out stream.