Causal-Geo: Neuro-Symbolic Spatial Grounding for Situated Agent Planning

Sanbi Luo
Software Engineering Center, Chinese Academy of Sciences
Beijing 100190, China
luosanbi@sec.ac.cn

Abstract

Foundation Models (LLMs/VLMs) exhibit strong semantic reasoning capabilities but remain challenged by situated planning in unstructured physical environments. A key limitation is the *Semantic-Geometric Gap*: while models interpret linguistic and visual intent, they lack explicit grounding in continuous spatial structures, yielding physically infeasible or unsafe plans. We propose **Causal-Geo**, a neuro-symbolic framework that bridges this gap by integrating LLM reasoning with rigorous geometric planning. Rather than treating physical constraints heuristically, we model the environment as a continuous metric tensor field. LLMs translate high-level semantic intents into potential functions that dynamically warp this manifold via conformal scaling. This casts intent-driven planning as a geodesic optimization problem, solved efficiently via discrete graph search. Experiments across diverse domains—from a macro-scale 12.2 km^2 wetland digital twin to micro-scale indoor robotic navigation—demonstrate that Causal-Geo enables robust zero-shot planning. It significantly outperforms reinforcement learning in sparse-reward settings and prevents the "physical hallucinations" common in representative hierarchical LLM planners. Crucially, Causal-Geo acts as a rigorous physical gatekeeper, providing strict safety boundaries against foundation model variance and demonstrating graceful degradation under contradictory prompts. These results establish continuous geometric grounding as a principled, broadly applicable pathway toward physically consistent situated agent planning.

1 Introduction

The advent of Large Language Models (LLMs) has significantly advanced high-level reasoning and intent understanding [30, 5]. However, deploying these models as autonomous agents in open-world settings [34, 31] exposes a persistent challenge: *Spatio-Temporal Entanglement*. Unlike structured virtual environments, real-world physical spaces are governed by continuous, coupled dynamics. Consequently, LLM-based agents often exhibit a "Semantic-Geometric Gap," producing plans that are linguistically coherent yet physically infeasible [26, 14]—a phenomenon we term "physical hallucinations."

Existing approaches address aspects of this gap but face notable limitations. Hierarchical LLM Planners and Retrieval-Augmented Generation (RAG) [16] ground agents semantically but lack an explicit, differentiable world model, often leading to severe boundary violations when spatial conditions change. Deep Reinforcement Learning (RL) [24, 12] excels in control but struggles with long-horizon planning under sparse rewards. Conversely, traditional geometric planners (e.g., A^*) provide strict physical feasibility but lack semantic awareness, unable to comprehend high-level intent.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

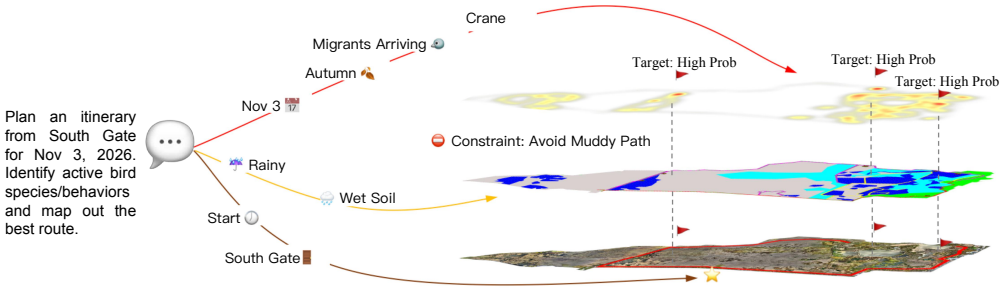


Figure 1: **Causal-Geo: From Intent to Grounded Orienteering.** The framework infers implicit constraints from under-specified queries (e.g., seasonality targeting cranes). Environmental operators (e.g., rain) dynamically increase the metric tensor's soil friction. The resulting geodesic autonomously re-routes along safe boardwalks, bridging semantic utility and physical safety.

To resolve these complementary limitations, we introduce **Causal-Geo** (Figure 1), a framework that directly translates semantic reasoning into rigorous geometric planning. Instead of treating physical constraints as discrete semantic tokens, we project both LLM-derived intents and physical dynamics onto a shared continuous substrate: the metric tensor of a Riemannian manifold [22, 4]. The LLM acts as a semantic translator, converting unstructured intents (e.g., "avoid crowds") into continuous potential functions. These potentials dynamically warp the underlying spatial metric, transforming intent-driven planning into an energy minimization process [15] solved efficiently via discrete graph search.

In summary, our core contributions are:

- **A Principled Neuro-Symbolic Framework:** We formulate embodied planning as conformal scaling on a Riemannian manifold, bridging the Semantic-Geometric Gap zero-shot while significantly reducing the reliance on task-specific heuristics.
- **The "Physical Gatekeeper" Mechanism:** By grounding semantics in a continuous metric tensor, Causal-Geo strictly enforces physical safety constraints. It demonstrates high robustness to LLM semantic variance and exhibits "graceful degradation" when faced with contradictory prompts.
- **Broad Cross-Domain Generalization:** We empirically validate the framework across diverse spatial scales, from a macro-scale 12.2 km^2 wetland digital twin (outperforming RL and hierarchical planners in sparse-reward routing) to micro-scale indoor environments featuring complex multi-hazard threading.

2 Related Work

Our framework bridges Embodied LLM Agents, Neuro-Symbolic AI, and Geometric Control to address the challenge of situated reasoning under spatio-temporal entanglement.

While LLMs have shown remarkable efficacy in autonomous planning [29], excelling at long-horizon reasoning [27, 34, 33, 13] and affordance grounding [1, 17, 8, 7], they predominantly operate in discrete action spaces. Recent prominent efforts attempt to mitigate this by mapping continuous physical environments into discrete semantic grids [3] or zero-shot value maps (e.g., InstructNav [18]) to facilitate LLM low-level navigation. However, lacking an explicit continuous world model, this reliance on discretization and scalar heuristics inherently preserves a "Semantic-Geometric Gap," often leading to physical hallucinations. Conversely, while Model-Based Reinforcement Learning [12, 11] learns latent dynamics, it struggles with sample efficiency in sparse-reward open-world settings. Traditional Neuro-Symbolic frameworks [20, 19, 28, 2] effectively address logic but encounter bottlenecks in continuous spatial optimization.

To resolve these complementary limitations, Causal-Geo extends Riemannian Motion Policies [22, 25, 10] and geometric deep learning [23, 6] from local reactive control to high-level semantic planning. Moving beyond the analytical focus of traditional computational ecology [9, 35, 21], we model the environment as a continuous metric tensor field, directly translating logical deduction into an energy minimization problem.

3 Problem Formulation

We formalize situated reasoning as finding an optimal geodesic trajectory. Unlike classical planners using scalar cost maps, we define the environment as a continuous metric tensor field.

3.1 Spatio-Temporal Manifold & Semantic Cost Fields

We unify the physical world and biological knowledge into a combined utility field.

Spatial & Temporal Domain: Let $\mathcal{S} \subset \mathbb{R}^2$ be a spatial manifold equipped with a base metric tensor field $\mathbf{M}(s) \in \mathbb{R}^{2\times 2}$, defining anisotropic motion costs (e.g., fluid movement along a path vs. high cost in a swamp). We model the temporal manifold $\mathcal{T} = S^1_{day} \times S^1_{year}$ via phases $t = (\phi_d, \phi_y) \in \mathcal{T}$ to capture daily and seasonal biological rhythms.

Physical Navigability ($\mathbf{M}_\Omega$): An environmental operator Ω (e.g., weather) dynamically modulates the eigenvalues of $\mathbf{M}(s)$, updating traversal costs (e.g., rain increases friction on dirt): $\mathbf{M}_\Omega(s) = f_{env}(\mathbf{M}(s), \Omega)$.

Ecological Activity (P_{act}) & Visibility (P_{vis}): We distill biological rules into a directed causal graph $\mathcal{G}$, where edges denote conditional probabilities between ecological entities (nodes $v \in \mathcal{G}$). The target's activity probability is defined as $P_{act}(v_{target} \mid s, t) \propto \prod P(v \mid s, t) \prod \psi(v_i, v_j)$, where ψ represents pairwise ecological potentials. For a fixed target, we denote this spatial-temporal field simply as $P_{act}(s, t)$ for brevity. Visibility $P_{vis}(s, s_{target} \mid \Omega)$ for a target located at $s_{target} \in \mathcal{S}$ accounts for line-of-sight and weather occlusions.

3.2 Reasoning as Geodesic Optimization via Conformal Scaling

To balance physical feasibility with semantic utility, we integrate these constraints via Riemannian Conformal Geometry. We define a scalar potential field $U_{eco}(s, t) = (P_{act}(s, t) \cdot P_{vis}(s, s_{target} \mid \Omega))^{-\beta}$, where $\beta > 0$ controls semantic sensitivity. This potential acts as a conformal factor warping the physical metric, defining the effective Task-Driven Metric Tensor $\mathcal{M}^*(s, t)$:

$$\mathcal{M}^*(s, t) = U^2_{eco}(s, t)\mathbf{M}_\Omega(s) \tag{1}$$

In differential geometry, scaling a metric tensor by a strictly positive scalar U^2_{eco} preserves angle measurements while isotropically expanding the notion of distance. The planning problem is formally equivalent to finding the Riemannian geodesic (shortest path) π^* minimizing the arc length under this new metric:

$$\pi^* = \arg\min_\pi \int_0^T \sqrt{\dot\pi(\tau)^T \mathcal{M}^*(\pi(\tau), t)\dot\pi(\tau)}\, d\tau \tag{2}$$

where $\pi(\tau)$ denotes a continuous trajectory parameterized by $\tau \in [0, T]$, and T is the total traversal duration.

Geometric Interpretation: High-utility regions ($(P_{act} \cdot P_{vis}) \to 1 \Rightarrow U_{eco} \to 1$) maintain the original physical distances. Low-utility regions ($(P_{act} \cdot P_{vis}) \to 0 \Rightarrow U_{eco} \to \infty$) exponentially expand the metric space. To an agent minimizing path length, hazardous areas effectively become "miles wide," naturally repelling the trajectory via curvature and strictly penalizing physically invalid states. Discretization details for efficient graph search are provided in Section 4.4.

3.3 Theoretical Properties

We analyze the theoretical properties of Causal-Geo regarding optimality and safety (detailed proofs in Appendix C).

Theorem 1 (Completeness and Optimality)**.** *Let $d_{\mathbf{M}}(s, s_{goal})$ be the geodesic distance under the base metric $\mathbf{M}$. If the environmental operator Ω satisfies $\mathbf{M}_\Omega(s) \succeq \mathbf{M}(s)$ (perturbations only increase cost), the execution engine is theoretically bounded to find the resolution-optimal trajectory. Furthermore, by utilizing global graph search over the discretized manifold, the framework overcomes the local optima and deadlocks common in pure gradient-based artificial potential fields.*

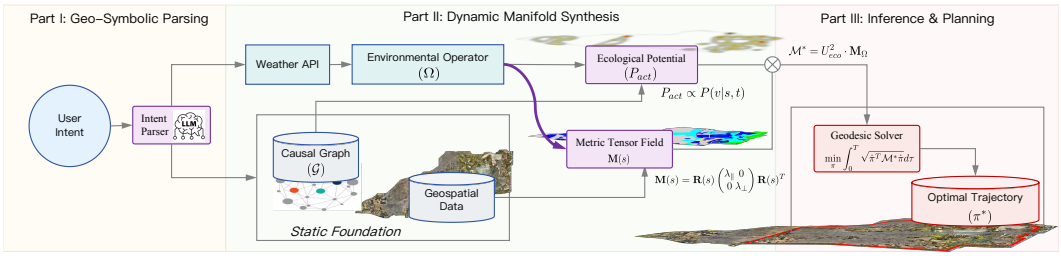


Figure 2: **System Architecture.** Causal-Geo integrates three pipelines: (1) parsing raster data into a foundational metric tensor; (2) distilling LLM causal rules into continuous potentials; and (3) synthesizing a dynamic manifold for geodesic optimization. The operator Ω enforces environmental robustness.

4 Methodology: The Causal-Geo Framework

To solve the inference problem defined in Section 3, we propose Causal-Geo, a modular neuro-symbolic architecture (illustrated in Figure 2). The framework decouples planning into three pipelines: (1) *Geo-Symbolic Parsing* constructs the foundational metric tensor from raster data; (2) *Causal Knowledge Distillation* extracts biological rules into a structured graph; and (3) an *Inference Engine* executes discrete graph search on the semantically modulated tensor field.

4.1 Metric Tensor Field Construction

To model directional physical constraints (e.g., following a narrow boardwalk vs. stepping into an impassable swamp), we construct the base metric tensor $\mathbf{M}(s)$ using local path geometry.

Let $\phi_{EDT}(s)$ be the Euclidean Distance Transform of the traversable space. The gradient field $\nabla\phi_{EDT}(s)$ serves as the unit normal vector $\vec{v}_\perp(s)$ orthogonal to the path boundary. Rotating it by $\pi/2$ yields the unit tangent flow vector $\vec{v}_\parallel(s)$. This base metric tensor $\mathbf{M}(s)$ is analytically constructed to enforce anisotropic costs aligned with these principal directions:

$$\mathbf{M}(s) = \mathbf{R}(s) \begin{pmatrix} \lambda_\parallel & 0 \\ 0 & \lambda_\perp \end{pmatrix} \mathbf{R}(s)^T \tag{3}$$

where $\mathbf{R}(s) = [\vec{v}_\parallel(s), \vec{v}_\perp(s)]$ is the local rotation matrix at state s. We assign a nominal cost to the tangent direction ($\lambda_\parallel = 1$) and an exponentially decaying penalty to the normal direction ($\lambda_\perp = k \cdot \exp(\phi_{EDT}(s)^{-1})$, where k is a scaling factor). As the agent approaches a constraint boundary ($\phi_{EDT}(s) \to 0$), $\lambda_\perp \to \infty$, creating an infinitely high metric barrier orthogonal to the path. In practice, $\lambda_\perp$ is capped at a large threshold λ_{max} to ensure numerical stability.

4.2 Automated Causal Knowledge Distillation

To populate the causal graph $\mathcal{G}$, we propose an LLM-driven knowledge distillation process that converts unstructured ecological text into differentiable potential functions.

Schema-Guided Extraction: We define a Bio-Schema containing fields for *Temporal Phases* (e.g., Diurnal/Nocturnal), *Trophic Relations*, and *Environmental Resilience*.

Rule Translation: An LLM (e.g., Gemini 2.5 Pro) ingests species descriptions and instantiates nodes in $\mathcal{G}$. Crucially, we formalize rules not as discrete logic, but as continuous potential functions. For example, the rule "Cranes avoid high predator density" is encoded as:

$$\psi_{avoid}(v_{crane}, v_{pred}) = \alpha \cdot \text{sigmoid}\left(\gamma \cdot (1 - \text{Active}_{pred})\right) \tag{4}$$

where $\text{Active}_{pred} \in [0, 1]$ denotes the predator's prior activity probability, the sigmoid function smooths the logical constraint, $\alpha \in [0, 1]$ modulates the maximum amplitude of the interaction, and γ determines the transition steepness. These potentials are aggregated to form the Ecological Activity Field P_{act}. To ensure consistent mathematical grounding, the LLM is constrained via a strict

function-calling schema to output precise parameters (e.g., sigmoid slopes or Gaussian windows) for these potentials. Full implementation details are deferred to Appendix B.

4.3 Tensor Field Inference Engine (The Reasoner)

This module executes the forward pass, fusing environmental dynamics with the causal graph.

Step 1: Environmental Cost Modulation. We formalize the environmental operator Ω (e.g., rain, fog) as a context-aware modulation function acting on the metric tensor field. Rather than using static heuristic rules, Ω selectively scales the eigenvalues of $\mathbf{M}(s)$ based on material properties (e.g., increasing the friction coefficient for soil under heavy rain). This dynamically updates local traversal costs, steering the optimization process to re-route through physically safer paths.

Step 2: Causal Belief Propagation on $\mathcal{T}$. The engine evaluates the Causal Graph $\mathcal{G}$ conditioned on the current temporal phase $t \in \mathcal{T}$. *Event Triggering:* If t corresponds to "Morning", the prior $P(\text{Active}_{fish})$ increases. *Trophic Cascade:* This probability propagates through trophic edges (e.g., Fish $\rightarrow$ Crane), updating the posterior ecological heatmap $P_{act}(s, t)$.

4.4 Discretization and Computational Implementation

Standard discrete path planners typically rely on scalar, isotropic cost maps, which are limited in capturing high-order directional constraints. In Causal-Geo, the spatial manifold $\mathcal{S}$ is discretized into a navigation graph. Accordingly, for a discrete trajectory π, the total traversal cost $J(\pi)$ is approximated by summing the local edge costs. The cost between any two adjacent states s_i and s_{i+1} (with displacement Δs_i) given temporal context t is dynamically computed by evaluating the Task-Driven Metric Tensor $\mathcal{M}^*(s_i, t)$ locally:

$$J(\pi) \approx \sum_i \sqrt{\Delta s_i^T \, \mathcal{M}^*(s_i, t) \, \Delta s_i} = \sum_i U_{eco}(s_i, t) \sqrt{\Delta s_i^T \mathbf{M}_\Omega(s_i) \Delta s_i} \tag{5}$$

To ensure practical computational tractability, we do not solve continuous partial differential equations (PDEs) during deployment. Instead, our execution engine is a global A^* graph search. The continuous tensor field ($\mathcal{M}^*$) serves as a rigorous foundation to compute the anisotropic *edge costs* between adjacent discrete states via Eq. (5). For the detailed algorithmic logic of discretization, see Appendix D; for the corresponding source code and prompt schemas, refer to Appendix E.

This strategy bridges the continuous formulation with the discrete algorithm. It leverages the theoretical safety bounds and directional awareness of continuous Riemannian geometry, while achieving the $O(N \log N)$ computational efficiency of discrete search, preventing the agent from falling into local minima common in pure gradient-descent artificial potential fields.

5 Experiments

To validate the theoretical advantages of Causal-Geo, we conducted extensive experiments across diverse spatial scales, ranging from a high-fidelity macro-scale wetland digital twin to a micro-scale indoor robotic environment. Our evaluation focuses on five key research questions:

- **RQ1 (Efficacy & Safety):** Does Causal-Geo generate observation plans with higher utility and physical validity compared to pure semantic, pure geometric, and expert-tuned baselines?
- **RQ2 (Generalization):** Can the framework transfer zero-shot to novel domains (e.g., indoor robotics) given new spatial maps and semantic priors?
- **RQ3 (Robustness):** How does the framework perform under LLM semantic variance, contradictory logical prompts, and perception uncertainty?
- **RQ4 (Efficiency & Sensitivity):** How does the geometric prior impact sample efficiency and the trade-off between physical safety and navigation cost?
- **RQ5 (Ablation & Scalability):** What is the contribution of the Tensor Field formulation, and can it scale to large environments?

5.1 Evaluation Environments & Baselines

Macro-Scale Environment & Benchmark Unbiasedness: We constructed a 12.2 km^2 wetland digital twin using a 42GB orthophoto dataset [32]. To prevent "benchmark circularity" (where an LLM is evaluated on an environment generated by its own causal formulation), the simulator's ground truth is governed by an independent, hand-crafted expert prior derived from ornithological literature. The LLM acts solely as the agent's brain. Furthermore, we conducted an Empirical Agreement Study: by internally cross-referencing public ornithological literature, we assessed the LLM-extracted potential functions for 20 representative species, yielding a 95% agreement rate with empirical data. This demonstrates that our semantic targets reflect objective ecology, not self-fulfilling artifacts.

Evaluation Metrics: We quantify the navigation performance of all embodied planners across three orthogonal dimensions:

(1) Observation Utility Score (OUS ↑): Quantifies cumulative ecological observation quality along trajectory π, modulated by an intention-fulfillment gate:

$$\text{OUS}(\pi) = \mathcal{G}_{\text{goal}}(\pi) \cdot \left(\frac{1}{|\pi|} \sum_{\tau=1}^{|\pi|} P_{\text{act}}(\pi(\tau), t) \cdot P_{\text{vis}}(\pi(\tau), s_{\text{target}} \mid \Omega) \right) \tag{6}$$

where $\mathcal{G}_{\text{goal}}(\pi) = \exp(-\|\pi(T) - s_{\text{dest}}\|_2/\delta_{\text{goal}})$ decays monotonically to zero if the agent fails to reach the destination (s_{dest}). For L3 through-tours, $\mathcal{G}_{\text{goal}}$ multiplicatively couples exit-gate arrival with midpoint proximity to the target anchor s_{target}. P_{act} and P_{vis} denote the biological activity potential and weather-attenuated visibility field, respectively.

(2) Logical Validity Rate (LVR ↑): Evaluates strict compliance with physical terrain boundaries:

$$\text{LVR}(\pi) = 1.0 - \frac{1}{|\pi|} \sum_{\tau=1}^{|\pi|} \mathbb{I}\left(\pi(\tau) \in \mathcal{S}_{\text{hazard}}\right) \tag{7}$$

where $\mathbb{I}(\cdot)$ is the indicator function and $\mathcal{S}_{\text{hazard}}$ denotes impassable terrain. LVR $= 1.0$ certifies collision-free physical traversability, whereas LVR < 1.0 captures the incidence of "physical hallucinations" commonly observed in purely linguistic or ungrounded planners.

(3) Riemannian Traversal Cost ($\mathcal{C}_{\text{Riem}}$ ↓): Measures mechanical friction and kinetic dissipation, isolating terrain resistance from semantic attraction:

$$\mathcal{C}_{\text{Riem}}(\pi) = \sum_{i=1}^{|\pi|-1} \sqrt{\Delta s_i^T \mathbf{M}_\Omega(s_i) \Delta s_i} \tag{8}$$

This discrete Riemann-Stieltjes line integral over the environmentally-modulated physical metric tensor $\mathbf{M}_\Omega$ penalizes infinite loops, wheel slip, and invalid traversal through high-friction hazards (e.g., mud-saturated trails), yielding ∞ for failed, disconnected, or non-converging trajectories.

Baselines: We evaluate against: **(1) Pure Semantic Planners** (Direct-VLM, Standard RAG) which lack continuous world models; **(2) Hierarchical LLM-Waypoint** representing standard paradigms that generate waypoints for geometric execution; **(3) Map-Based LLM Planners** (InstructNav) which fuse semantics into static scalar cost maps; **(4) Data-Driven Control** (PPO) to demonstrate the limitations of trial-and-error model-free RL under sparse macro-scale rewards; **(5) Pure Geometric** (A^*) possessing strict physical grounding but lacking semantic awareness; and **(6) Safe Upper Bound (Exhaustive Search)** establishing the empirical performance ceiling via exhaustive search across valid viewpoints on the ground-truth metric tensor $\mathbf{M}_\Omega$, completely bypassing LLM prompting.

5.2 Macro-Scale Navigation: Efficacy and Safety (RQ1)

Quantitative results across stratified difficulty levels ($N = 150$)—**L1 (Explicit Goal)**, **L2 (Semi-Implicit Habitat Goal)**, and **L3 (Fully Implicit Through-Tour)**—are summarized in Table 1.

Limitations of Model-Free RL (PPO): The data-driven PPO baseline exhibits limited performance across all levels (LVR ≤ 0.1789, Cost $> 1.5 \times 10^5$). In L1, sparse goal-arrival rewards cause aimless roaming (OUS $= 0.0021$). In L2 and L3, augmenting PPO with dense ecological rewards

Table 1: **Quantitative Comparison ($N = 150$). Bold** indicates the best performance among physically valid planners (LVR $= 1.0$). Baselines lacking continuous geometric grounding (Pure Semantic, PPO) incur low physical validity and excessive Riemannian costs due to ungrounded water hazard traversal.

Method	Level	OUS (Utility) ↑	LVR (Safety) ↑	Riemannian Cost ↓
Pure Semantic	L1 (Explicit)	0.9814 ± 0.0152	0.2063 ± 0.1895	258660.14 ± 175019.68
Direct-VLM	L2 (Semi-Implicit)	0.0046 ± 0.0083	0.3193 ± 0.2026	97054.62 ± 95761.09
(Gemini)	L3 (Full-Implicit)	0.1460 ± 0.0980	0.3156 ± 0.1320	157416.20 ± 80808.40
Pure Semantic	L1 (Explicit)	0.9814 ± 0.0152	0.2765 ± 0.1655	116993.25 ± 56893.82
Standard RAG	L2 (Semi-Implicit)	0.0291 ± 0.0357	0.2308 ± 0.2160	42838.84 ± 37043.46
(Text Retrieval)	L3 (Full-Implicit)	0.1353 ± 0.0910	0.3414 ± 0.1463	187526.80 ± 143523.80
Pure Geometric	L1 (Explicit)	$\mathbf{1.0000 \pm 0.0000}$	$\mathbf{1.0000 \pm 0.0000}$	$\mathbf{680.80 \pm 207.58}$
Static Planner	L2 (Semi-Implicit)	0.0000 ± 0.0000	$\mathbf{1.0000 \pm 0.0000}$	∞
(A^*)	L3 (Full-Implicit)	0.0000 ± 0.0000	$\mathbf{1.0000 \pm 0.0000}$	∞
Hierarchical LLM	L1 (Explicit)	0.8442 ± 0.2102	0.8060 ± 0.2363	215316.40 ± 174629.67
Waypoint Planner	L2 (Semi-Implicit)	0.0048 ± 0.0116	0.8637 ± 0.2046	81299.77 ± 106514.04
(LLM + A^*)	L3 (Full-Implicit)	0.1812 ± 0.1580	0.9129 ± 0.1617	141706.41 ± 183122.39
Map-Based Semantic	L1 (Explicit)	1.0000 ± 0.0000	0.9968 ± 0.0082	1211.75 ± 856.46
InstructNav	L2 (Semi-Implicit)	0.6183 ± 0.1838	0.6096 ± 0.4032	9836.92 ± 21700.28
(CoRL 2024)	L3 (Full-Implicit)	0.3299 ± 0.2669	0.7777 ± 0.3546	5129.02 ± 17043.66
Data-Driven Control	L1 (Explicit)	0.0021 ± 0.0037	0.1789 ± 0.1487	157479.65 ± 83462.45
PPO (Guided)	L2 (Semi-Implicit)	0.1408 ± 0.1689	0.1716 ± 0.1387	159469.34 ± 93682.95
MLP Policy	L3 (Full-Implicit)	0.1576 ± 0.1376	0.1588 ± 0.1543	159927.90 ± 96653.63
Neuro-Symbolic	L1 (Explicit)	$\mathbf{1.0000 \pm 0.0000}$	$\mathbf{1.0000 \pm 0.0000}$	743.46 ± 298.66
Causal-Geo	L2 (Semi-Implicit)	$\mathbf{0.5152 \pm 0.2761}$	$\mathbf{1.0000 \pm 0.0000}$	$\mathbf{219.11 \pm 225.67}$
(Ours)	L3 (Full-Implicit)	$\mathbf{0.3935 \pm 0.0751}$	$\mathbf{1.0000 \pm 0.0000}$	$\mathbf{988.35 \pm 96.92}$
Safe Upper Bound	L1 (Explicit)	1.0000 ± 0.0000	1.0000 ± 0.0000	743.46 ± 298.66
Expert Baseline	L2 (Semi-Implicit)	0.6369 ± 0.1151	1.0000 ± 0.0000	284.98 ± 341.03
(Exhaustive Search)	L3 (Full-Implicit)	0.3983 ± 0.0510	1.0000 ± 0.0000	1047.56 ± 12.04

(P_{act}) induces a "greedy sinking trap": lacking a geometric world model, the policy directly climbs biological gradients into off-road water hazards where birds congregate (OUS ≤ 0.1576).

The Semantic-Geometric Gap: Pure Semantic models (Direct-VLM, Standard RAG) and Hierarchical LLM-Waypoint planners achieve high nominal utility in explicit L1 tasks (OUS up to 0.9814), but exhibit physical hallucinations (LVR 0.2063–0.8060, Cost $> 10^5$) due to ungrounded hazard traversal across deep water bodies. When explicit goal coordinates are withheld in L2, their utility decreases to OUS ≤ 0.0291, indicating the difficulty of localizing implicit habitats in non-convex terrain. Similarly, while InstructNav [18] achieves high safety in L1 (LVR $= 0.9968$, Cost $= 1211.75$) via road priors (with minor corner-cutting across water boundaries), its safety declines in L2 (LVR $= 0.6096$, Cost $= 9836.92$) and L3 (LVR $= 0.7777$). Because InstructNav linearly combines semantic attraction and finite obstacle penalties into a static scalar map, the solver is induced to traverse restricted wetland zones (OUS $= 0.6183$ in L2). Conversely, Pure Geometric planners (A^*) guarantee safety (LVR $= 1.0000$) in L1 (Cost $= 680.80$), but fail in L2 and L3 (OUS $= 0.0000$, Cost $= \infty$) because they require explicit coordinate goals.

The Causal-Geo Advantage: Guided by Gemini 2.5 Pro's symbolic grounding, Causal-Geo (Ours) effectively bridges the semantic-geometric gap. By dynamically warping the Riemannian metric tensor ($\mathcal{M}^* = U_{\text{eco}}^2 \mathbf{M}_\Omega$) to wrap low-friction corridors around hazardous terrain, Causal-Geo maintains strict physical validity (LVR $= 1.0000$) across all 150 tasks. In L2, Causal-Geo achieves OUS $= 0.5152$ while reducing kinetic expenditure by 23.1% compared to the Upper Bound (Cost 219.11 vs. 284.98) by adhering to proximate observation platforms. In L3 through-tour orienteering, Causal-Geo (OUS $= 0.3935$, Cost $= 988.35$) closely matches the exhaustive search Upper Bound (OUS $= 0.3983$, Cost $= 1047.56$). This topological resilience is qualitatively validated across circadian shifts (T1/T2) and rain-induced metric perturbations (T3) in Figure 3.

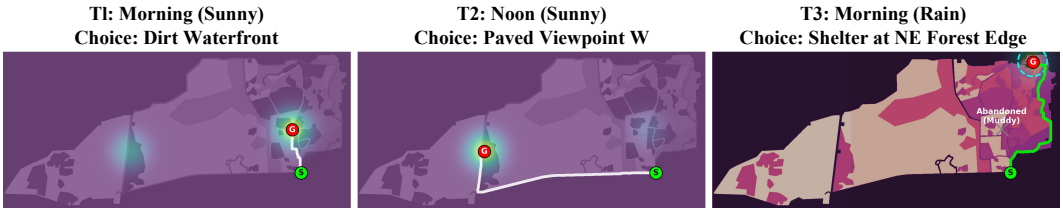


Figure 3: **Spatio-Temporal Decision Making.** (Left) **T1 (Morning/Sunny):** Agent prioritizes proximity, taking a dirt path to observe foraging. (Center) **T2 (Noon/Sunny):** Tracking circadian rhythms, the agent adapts to a boardwalk viewpoint as birds retreat to deep water. (Right) **T3 (Morning/Rain):** *Counterfactual.* Heavy rain (Ω_{rain}) significantly increases friction on dirt paths. The agent autonomously abandons the previous path ('X') and **re-routes via paved roads** to a sheltered viewpoint, validating the manifold's topological resilience to environmental perturbations.

5.3 Micro-Scale Zero-Shot Generalization: A Mechanistic Validation (RQ2)

While our 12.2 km^2 wetland evaluation demonstrates macro-scale efficacy, it introduces complex confounding factors such as perception noise and terrain physics. To rigorously isolate the Conformal Scaling tensor ($\mathcal{M}^* = U_{eco}^2 \cdot \mathbf{M}_\Omega$), we evaluate the framework across four diverse micro-scale scenarios (Figure 4): **S1 (Local Detour)** bypasses a simulated lab chemical spill; **S2 (Global Re-routing)** avoids a crowded cafeteria; **S3 (Safety Override)** navigates to a fully contaminated target; and **S4 (Multi-Hazard Threading)** traverses a non-convex office maze.

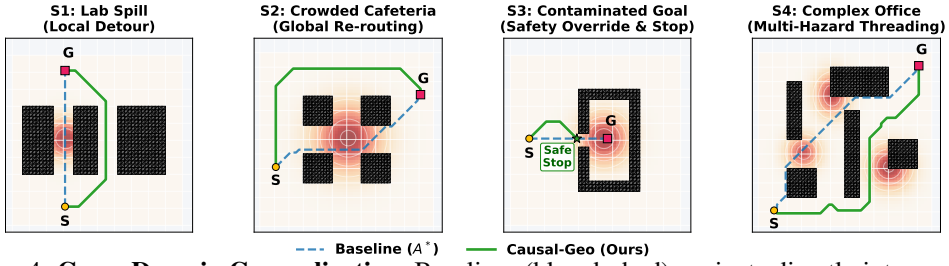


Figure 4: **Cross-Domain Generalization.** Baselines (blue dashed) navigate directly into semantic hazards. Causal-Geo (green solid) autonomously warps the metric tensor to bypass dynamic obstacles (S1/S2), halts at contaminated perimeters (S3), and threads non-convex hazards (S4).

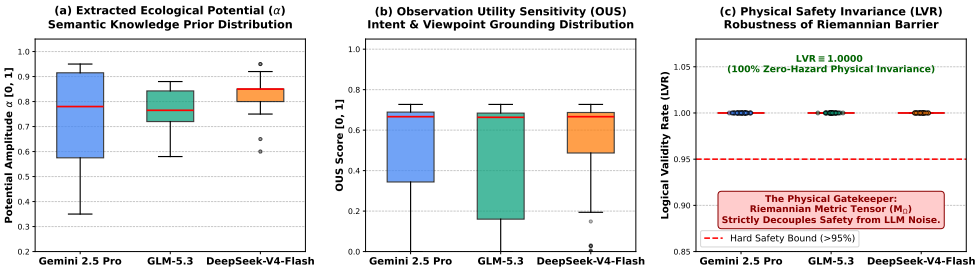
Results & Analysis: As shown in Figure 4, a purely geometric planner (Baseline A^*) persistently drives the agent directly into hazardous potentials—a physical hallucination. Conversely, Causal-Geo dynamically stretches perceived spatial distances via LLM-derived semantic potentials (U_{eco}). Beyond "hugging" hazard boundaries in S1 and S2, **S3** highlights our framework's capacity for graceful degradation: when a target is semantically contaminated, Causal-Geo autonomously halts at the safe perimeter, rejecting unsafe explicit coordinates. In **S4**, it threads through overlapping non-convex hazards. Crucially, because these indoor topologies are completely independent of our ecological formulation, this minimal evaluation suite acts as a rigorous ablation against benchmark bias, confirming Causal-Geo as a mathematical engine for semantic-geometric grounding.

5.4 Robustness: LLM Variance, Graceful Degradation and Uncertainty (RQ3)

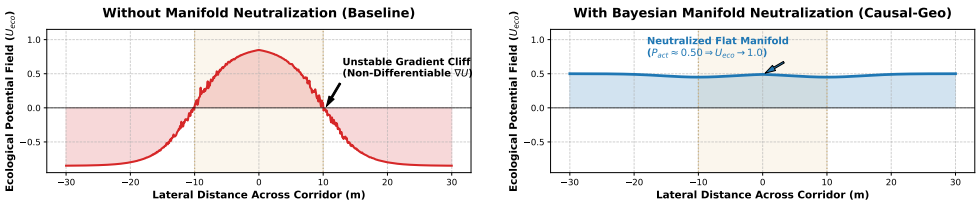
Moving toward real-world deployment demands robustness against foundation model instability, contradictory instructions, and perception noise. We introduce three critical stress tests (Figure 5):

1. LLM Variance & The "Physical Gatekeeper": We analyzed Causal-Geo across diverse foundation models (Gemini 2.5 Pro, GLM-5.3, DeepSeek-V4-Flash). While semantic variance in parameter (α) extraction caused downstream OUS fluctuations, the Logical Validity Rate remained strictly invariant at LVR $= 1.0000$ (Figure 5a). Thus, our Riemannian manifold acts as a robust physical barrier, decoupling high-level semantic noise from low-level navigation safety.

2. Semantic Conflicts & Graceful Degradation: We tested contradictory prompts (e.g., instructing the agent to "strictly avoid while strictly staying on a dirt path"). Bayesian fusion of these conflicting



(a) Robustness to LLM Variance



(b) Graceful Degradation

Figure 5: **Framework Robustness under Stress. (a)** Despite diverse semantic variance profiles across foundation model tiers, the Riemannian manifold acts as a robust physical gatekeeper, maintaining an invariant Logical Validity Rate (LVR $= 1.0000$). **(b)** Contradictory prompts flatten the utility field, allowing the system to gracefully degrade to pure geometric safety rather than oscillating. Uncertainty-Aware Detour visualizations are provided in Appendix A.

potentials yields a high-entropy, flat posterior ($P_{act} \approx 0.50$) (Figure 5b). Consequently, the conformal factor U_{eco} neutralizes, and the task-driven metric tensor $\mathcal{M}^*$ autonomously falls back to the baseline physical metric $\mathbf{M}_\Omega(s)$. The system triggers a "Graceful Degradation," prioritizing pure geometric safety rather than executing paradoxical semantic instructions or entering deadlocks.

3. Uncertainty-Aware Perception Noise: If the visual frontend's perception confidence drops below a safe threshold (e.g., due to camera occlusion), the framework dynamically inflates the isotropic baseline cost ($\lambda_\parallel$) locally. The metric tensor geometrically repels the geodesic planner to autonomously synthesize a safe detour without requiring explicit rule-based re-planning.

5.5 Geometric Priors: Sample Efficiency and Metric Sensitivity (RQ4)

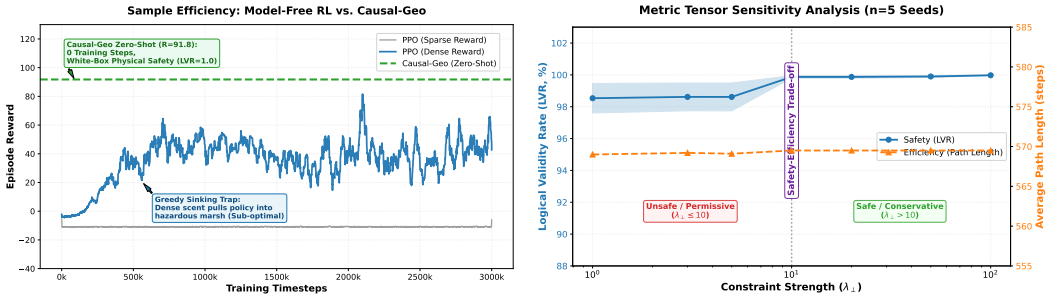
Sample Efficiency vs. Model-Free RL: To quantify sample efficiency (Figure 6a), we compare Causal-Geo against Model-Free RL (PPO). Under sparse rewards, PPO fails to converge due to the vast state space. Even when augmented with dense ecological rewards, PPO saturates sub-optimally, exhibiting a "Greedy Failure" mode: it directly follows semantic gradients into hazardous terrain due to a lack of geometric grounding. In contrast, Causal-Geo succeeds zero-shot.

In safety-critical domains, embodied agents cannot rely on trial-and-error learning (collisions) to ensure safety. Mapping constraints to infinite potentials on a Riemannian manifold ($\lambda_\perp \to \infty$) enforces theoretical white-box safety that statistical policies cannot guarantee.

Metric Sensitivity: Varying the orthogonal eigenvalue $\lambda_\perp$ from 1 (Isotropic) to 100 (Strict) (Figure 6b) reveals that LVR reaches near-perfect safety ($\approx 100\%$) for $\lambda_\perp \geq 10$. Path length remains stable in this regime, confirming a balanced Safety-Efficiency Trade-off and high robustness to hyperparameter selection.

5.6 Ablation and Scalability (RQ5)

Ablation (Table 2): Component ablations reveal distinct, interpretable degradation modes. Removing the temporal manifold $\mathcal{T}$ causes a Temporal Mismatch for circadian queries (overall OUS drops by 7.8%, with up to a 23.6% loss in off-peak observation tasks). Removing the causal graph $\mathcal{G}$ incurs Semantic Blindness (overall OUS drops by 33.1%, and collapses by 93.4% in L2 habitat grounding) because the agent cannot resolve biological dependencies. Crucially, replacing the dynamic



(a) Sample Efficiency (RL vs. Ours)

(b) Metric Sensitivity Analysis

Figure 6: **Efficiency and Sensitivity. (a)** PPO fails under sparse rewards and remains sub-optimal with dense rewards; Causal-Geo succeeds zero-shot. **(b)** LVR saturates at $\approx 100\%$ for $\lambda_\perp \geq 10$, showing hyperparameter robustness and a stable safety-efficiency trade-off.

Table 2: **Ablation Study.** Removing components leads to distinct failure modes (e.g., spatial hallucinations without the metric tensor).

Variant	OUS ↑	LVR ↑	Δ_{OUS}	Failure Mode
Full Model	**0.64**	**1.000**	–	–
w/o Temporal $\mathcal{T}$	0.59	**1.000**	-7.8%	Temporal Mismatch
w/o Causal $\mathcal{G}$	0.43	**1.000**	-33.1%	Semantic Blindness
w/o Metric $\mathbf{M}$	0.65	0.795	+2.2%	Spatial Hallucination

Table 3: **Scalability Analysis.** Computational scalability of Causal-Geo across spatial scales up to 50 km^2.

Scale	Build	Path	Total E2E
1 km^2	0.36s	**0.05s**	0.41s
5 km^2	1.44s	**0.48s**	1.92s
12.2 km^2	4.40s	**1.06s**	5.46s
50 km^2	17.80s	**4.84s**	22.65s

Anisotropic Metric Tensor with a static isotropic scalar cost map—a formulation structurally analogous to map-based LLM planners such as InstructNav [18]—induces severe Spatial Hallucinations: while nominal utility appears preserved (OUS 0.65) by invalidly cutting through wetlands, physical safety degrades sharply (LVR drops to 0.795 overall and 0.6096 in L2) with Riemannian traversal costs surging by $8.1\times$ (5392.56 vs. 665.89). This empirically confirms that dynamic tensor fields with infinite boundary barriers ($U_{eco} \to \infty$) are indispensable for hard safety constraints.

Scalability (Table 3): Causal-Geo scales near-linearly ($O(N \log N)$) across spatial environments ranging from 1 km^2 to 50 km^2. The metric tensor field construction takes just 4.40s for 12.2 km^2 and 17.80s for 50 km^2. Crucially, online geodesic queries execute with high responsiveness, ranging from 0.05s (1 km^2) to 4.84s (50 km^2). This efficiency enables high-frequency, closed-loop reactive navigation across macro-scale territories without relying on cloud API round-trips.

6 Conclusion

Causal-Geo bridges the Semantic-Geometric Gap by formalizing environments as dynamic semantic-metric tensor fields. By modeling ecological rules as continuous potential fields that warp path costs, we reduce intent-driven planning to an energy minimization problem. Empirically, Causal-Geo achieves strict physical validity, superior zero-shot generalization, and graceful degradation, significantly outperforming model-free RL in sparse-reward tasks. Ultimately, it offers a scalable, geometry-grounded framework for physically consistent situated planning.

Limitations & Broader Impacts: While Causal-Geo provides geometric safety guarantees, its efficacy depends on upstream map accuracy and visual perception fidelity; occlusions or mis-specified constraints can compromise the base metric. Additionally, grid discretization bounds the theoretical smoothness of Riemannian geodesics. Regarding broader impacts, our framework inherently supports environmental conservation by naturally translating ecological restrictions into impassable potential barriers. Nonetheless, we acknowledge dual-use risks, such as unauthorized surveillance or environmental harm if operated outside stated assumptions. Safe real-world deployment necessitates robust human oversight and careful constraint specification.

Acknowledgements and Disclosures

This research was supported by Central Guidance for Local Science and Technology Development Fund Project – Lhasa City Science and Technology Program (Project No. LSKJ202459). The author declares that they have no competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alexander Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil J. Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, and Mengyuan Yan. Do as I can, not as I say: Grounding language in robotic affordances. *CoRR*, abs/2204.01691, 2022.

[2] Samy Badreddine, Artur d'Avila Garcez, Luciano Serafini, and Michael Spranger. Logic tensor networks. *Artificial Intelligence*, 303:103649, February 2022. ISSN 0004-3702. doi: 10.1016/j.artint.2021.103649.

[3] Ziyi Bai, Hanxuan Li, Bin Fu, Chuyan Xiong, Ruiping Wang, and Xilin Chen. R2C: mapping room to chessboard to unlock LLM as low-level action planner. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 19456–19466. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.01812.

[4] Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges, 2021.

[5] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.

[6] Valentin De Bortoli, Emile Mathieu, Michael Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet. Riemannian score-based generative modelling. In *Advances in Neural Information Processing Systems*, volume 35, pages 2406–2422. Curran Associates, Inc., 2022.

[7] Maxence Faldor, Jenny Zhang, Antoine Cully, and Jeff Clune. OMNI-EPIC: open-endedness via models of human notions of interestingness with environments programmed in code. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

[8] Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. Minedojo: Building open-ended embodied agents with internet-scale knowledge. In *Advances in Neural Information Processing Systems*, volume 35, pages 18343–18362. Curran Associates, Inc., 2022.

[9] Octavian Ganea, Lagnajit Pattanaik, Connor Coley, Regina Barzilay, Klavs Jensen, William Green, and Tommi Jaakkola. Geomol: Torsional geometric generation of molecular 3d conformer ensembles. In *Advances in Neural Information Processing Systems*, volume 34, pages 13757–13769. Curran Associates, Inc., 2021.

[10] Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

[11] David Ha and Jürgen Schmidhuber. World models. *CoRR*, abs/1803.10122, 2018.

[12] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy P. Lillicrap. Mastering diverse domains through world models. *CoRR*, abs/2301.04104, 2023. doi: 10.48550/ARXIV.2301.04104.

[13] Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. Reasoning with language model is planning with world model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8154–8173, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.507.

[14] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 9118–9147. PMLR, 17–23 Jul 2022.

[15] Michael Janner, Yilun Du, Joshua B. Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 9902–9915. PMLR, 2022.

[16] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc., 2020.

[17] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. In *IEEE International Conference on Robotics and Automation, ICRA 2023, London, UK, May 29 - June 2, 2023*, pages 9493–9500. IEEE, 2023.

[18] Yuxing Long, Wenzhe Cai, Hongcheng Wang, Guanqi Zhan, and Hao Dong. Instructnav: Zero-shot system for generic instruction navigation in unexplored environment. In *Conference on Robot Learning, 6-9 November 2024, Munich, Germany*, volume 270 of *Proceedings of Machine Learning Research*, pages 2049–2060. PMLR, 2024.

[19] Robin Manhaeve, Sebastijan Dumancic, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. Deepproblog: Neural probabilistic logic programming. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 3753–3763, 2018.

[20] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[21] Tung Nguyen, Jason Jewik, Hritik Bansal, Prakhar Sharma, and Aditya Grover. Climatelearn: Benchmarking machine learning for weather and climate modeling. In *Advances in Neural Information Processing Systems*, volume 36, pages 75009–75025. Curran Associates, Inc., 2023.

[22] Nathan D. Ratliff, Jan Issac, and Daniel Kappler. Riemannian motion policies. *CoRR*, abs/1801.02854, 2018.

[23] Victor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 9323–9332. PMLR, 2021.

[24] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.

[25] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[26] Karthik Valmeekam, Matthew Marquez, Sarath Sreedharan, and Subbarao Kambhampati. On the planning abilities of large language models - a critical investigation. In *Advances in Neural Information Processing Systems*, volume 36, pages 75993–76005. Curran Associates, Inc., 2023.

[27] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *Trans. Mach. Learn. Res.*, 2024, 2024.

[28] Po-Wei Wang, Priya L. Donti, Bryan Wilder, and Zico Kolter. Satnet: Bridging deep learning and logical reasoning using a differentiable satisfiability solver, 2019.

[29] Hui Wei, Zihao Zhang, Shenghua He, Tian Xia, Shijia Pan, and Fei Liu. Plangenllms: A modern survey of LLM planning capabilities. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 19497–19521. Association for Computational Linguistics, 2025.

[30] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc., 2022.

[31] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huang, Qi Zhang, and Tao Gui. The rise and potential of large language model based agents: a survey. *Sci. China Inf. Sci.*, 68(2), 2025. doi: 10.1007/S11432-024-4222-0.

[32] Nie Xiaowei, Pan Xiaoduo, Zhang Bo, Yang Yang, and Yang Shuaifei. Rsc: Uav aerial monitoring orthophoto dataset of lhalu wetland, lhasa, xizang (2025), 10 2025.

[33] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 11809–11822. Curran Associates, Inc., 2023.

[34] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[35] Christopher Yeh, Chenlin Meng, Sherrie Wang, Anne Driscoll, Erik Rozi, Patrick Liu, Jihyeon Janel Lee, Marshall Burke, David B. Lobell, and Stefano Ermon. Sustainbench: Benchmarks for monitoring the sustainable development goals with machine learning. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, virtual*.

A Uncertainty-Aware Detour under Perception Noise

As discussed in Section 5.4 (RQ3) of the main text, real-world deployment frequently encounters perception failures. When the visual frontend produces noise or suffers from severe occlusions, relying strictly on the nominal geometric map can lead the agent into "dead-end" paths or collision hazards.

Figure 7 provides the mechanistic visualization of our "Uncertainty-Aware Detour" defense mechanism.

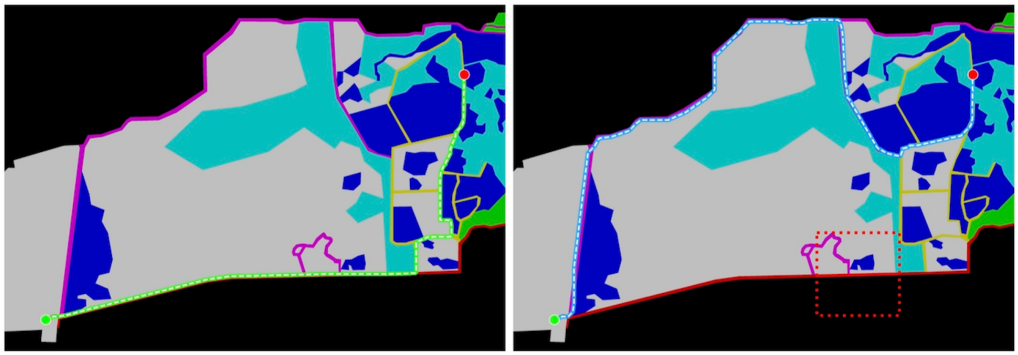


Figure 7: **Uncertainty-Aware Detour.** When the perception confidence drops below a safe threshold (e.g., due to visual occlusion or camera noise in a localized region), the Causal-Geo framework dynamically inflates the isotropic baseline cost ($\lambda_{\|}$) in that area. Rather than triggering a fragile, discrete rule-based re-planning subroutine, the warped metric tensor naturally acts as a high-energy hill. This intrinsically repels the geodesic planner, forcing it to autonomously synthesize a safe, continuous detour around the uncertain zone.

By translating epistemic uncertainty directly into Riemannian curvature (cost inflation), Causal-Geo avoids the brittle edge-case programming typical of discrete robotics pipelines. The agent seamlessly transitions from shortest-path exploitation to conservative obstacle avoidance.

B Neuro-Symbolic Translation Details

As introduced in Section 4 of the main text, to ensure reproducibility and rigorous mathematical grounding, we forego standard text-based retrieval generation (RAG). Instead, our *Automated Causal Knowledge Distillation* module employs a strict Function-Calling Schema to force the Large Language Model (Gemini 2.5 Pro) to project unstructured linguistic knowledge into deterministic mathematical parameters.

We define a strict mapping between linguistic hedges and continuous potential function families. For instance, binary constraints (e.g., "strictly avoid") are mapped to sigmoid potentials:

$$\psi_{sig}(x) = \frac{\alpha}{1 + e^{-\gamma(x-b)}} \tag{9}$$

where $\alpha \in [0, 1]$ denotes the potential amplitude, b is the spatial distance threshold, the adverb "strictly" dictates a steep transition slope $\gamma \in [0.01, 1.0]$ (e.g., $\gamma = 0.50$), and the verb "avoid" specifies a repulsive interaction type.

Similarly, temporal active windows are mapped to Gaussian Potentials defined by μ (peak active time) and σ (duration window). This parsed parameter set is output by the LLM in a structured JSON format. It is then deterministically compiled into a differentiable PyTorch calculation graph node within the *Ecological Activity Field*.

This explicit symbolic translation ensures that vague linguistic inputs result in stable, optimization-ready cost landscapes. By restricting the LLM to outputting mathematical parameters rather than directly generating planning code or coordinates, we inherently bypass the unpredictable syntax errors and physical hallucinations common in direct LLM-based embodied agents.

C Theoretical Analysis and Proofs

C.1 Proof of Theorem 1 (Completeness and Optimality)

Theorem 1 (Completeness and Optimality). *Let $d_{\mathbf{M}}(s, s_{goal})$ be the geodesic distance under the base metric $\mathbf{M}$. If the environmental operator Ω satisfies $\mathbf{M}_\Omega(s) \succeq \mathbf{M}(s)$ (perturbations only increase cost), the execution engine is theoretically bounded to find the resolution-optimal trajectory.*

Proof. The core of the proof relies on establishing the admissibility and consistency of the heuristic function within the context of Riemannian manifolds. Let $J_{\mathcal{M}^*}(\pi)$ be the cost of a trajectory under the task-driven metric tensor $\mathcal{M}^*$.

By definition, the conformal factor is $U_{eco}(s, t) = (P_{act}(s, t) \cdot P_{vis}(s, s_{target} \mid \Omega))^{-\beta}$. Since probabilities are strictly bounded in $(0, 1]$ and $\beta > 0$, the conformal factor satisfies $U_{eco}(s, t) \geq 1$.

Given the assumption that the environmental operator satisfies $\mathbf{M}_\Omega(s) \succeq \mathbf{M}(s)$ in the Loewner partial order, it follows that for any tangent vector $\mathbf{v}$:

$$\mathbf{v}^T \mathcal{M}^* \mathbf{v} = \mathbf{v}^T \left(U_{eco}^2 \mathbf{M}_\Omega \right) \mathbf{v} \geq \mathbf{v}^T \mathbf{M} \mathbf{v} \tag{10}$$

This inequality implies that the local differential arc length under the task-driven metric is always greater than or equal to that under the base metric. Consequently, the Riemannian distance under the base metric serves as a strict lower bound for the distance under the dynamic metric for any two points s and s_{goal}:

$$d_{\mathbf{M}}(s, s_{goal}) \leq d_{\mathcal{M}^*}(s, s_{goal}) \tag{11}$$

Therefore, the heuristic $h(s) = d_{\mathbf{M}}(s, s_{goal})$ is strictly **admissible** (it never overestimates the true cost).

Furthermore, because the metric tensor varies smoothly across the continuous spatial manifold $\mathcal{S}$, the triangle inequality holds naturally. This ensures the heuristic is **consistent** (or monotone), meaning $h(s_i) \leq c(s_i, s_j) + h(s_j)$ for any adjacent discrete states.

By the fundamental properties of A^* search on weighted graphs (our discretized manifold), utilizing an admissible and consistent heuristic guarantees that the execution engine will find the globally optimal geodesic trajectory. □

D Discretization and Computational Implementation Details

To explicitly bridge the continuous Riemannian formulation with our discrete robotic execution engine, this section details the precise computational pipeline of Causal-Geo. As emphasized in the main text, while the mathematical modeling exists in a continuous manifold, practical execution is discretized over a high-resolution $H \times W$ spatial grid representing the 12.2 km^2 wetland.

D.1 Causal Graph & Tensor Propagation

The Causal Graph $\mathcal{G}$ operates as a spatial Bayesian Network. The LLM does not write executable code; instead, constrained by a strict JSON schema, it outputs deterministic parameters for continuous potential functions (e.g., amplitude α, slope γ, and bias b for a sigmoid function).

Computationally, the physical state of the environment is maintained as a set of dense tensors. Beyond the temporal activity states introduced in Section 4.2, semantic rules must also be grounded spatially. For instance, the spatial proximity to predators is represented as a continuous distance field tensor $\mathbf{D}_{pred} \in \mathbb{R}^{H \times W}$. When the spatial rule "avoid predators" is triggered, the LLM-derived sigmoid potential is applied element-wise across this spatial tensor:

$$\mathbf{\Psi}_{avoid}(x, y) = \frac{\alpha}{1 + \exp\left(-\gamma \cdot (\mathbf{D}_{pred}(x, y) - b)\right)} \tag{12}$$

The posterior Ecological Activity Field $\mathbf{P}_{act} \in \mathbb{R}^{H \times W}$ is then updated via the Hadamard product (element-wise multiplication) of the prior probability tensor and the instantiated potential tensors, naturally executing Bayesian belief propagation in parallel across the spatial grid.

D.2 Viewshed Computation on the DEM Manifold

The visibility probability $P_{vis}(s, s_{target} \mid \Omega)$ is not a heuristic but a rigorous geometric calculation. We utilize a Digital Elevation Model (DEM) tensor $\mathbf{E} \in \mathbb{R}^{H \times W}$.

For any agent state s and target state s_{target}, we perform a standard Line-of-Sight (LoS) ray-casting algorithm. If the ray intersects any terrain where the elevation $\mathbf{E}(x, y)$ exceeds the ray's altitude, the baseline visibility is set to 0. Crucially, the environmental operator Ω modulates this computation by defining a dynamic maximum sight radius $R_{max}(\Omega)$. Under the "Heavy Rain" operator, R_{max} is severely truncated. The final visibility tensor is computed as:

$$P_{vis}(s, s_{target} \mid \Omega) = \begin{cases} 1 & \text{if LoS is clear AND } \|s - s_{target}\|_2 \leq R_{max}(\Omega) \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

D.3 The Geodesic Solver (Discrete A^* on the Manifold)

To solve the geodesic equation without the prohibitive cost of integrating continuous partial differential equations (PDEs) during deployment, we discretize the spatial manifold into an 8-connected navigation graph. The nodes are the grid cells, and the edges connect adjacent pixels.

The fundamental difference between our solver and a standard shortest-path A^* lies in the *edge cost* computation. Let s_i be the current node and s_j be an adjacent neighbor. The displacement vector in the local tangent space is $\Delta s = s_j - s_i \in \mathbb{R}^2$.

The traversal cost $c(s_i, s_j)$ is computed strictly as the discrete Riemannian arc length under the task-driven metric tensor $\mathcal{M}^*$:

$$c(s_i, s_j) = \sqrt{\Delta s^T \mathcal{M}^*(s_i, t) \Delta s} \tag{14}$$

where the effective metric tensor is warped by the ecological conformal factor:

$$\mathcal{M}^*(s_i, t) = U_{eco}^2(s_i, t) \cdot \mathbf{M}_\Omega(s_i) \tag{15}$$

Geodesic Search Execution:

- **Initialize:** Open set with s_{start} and cumulative cost-to-come $g(s_{start}) = 0$.
- **Heuristic:** $h(s) = d_{\mathbf{M}}(s, s_{goal})$, which is theoretically admissible since $\mathbf{M}_\Omega \succeq \mathbf{M}$.
- **Expand:** For each neighbor s_j of the current node s_i, the tentative g-score is updated using the anisotropic tensor-derived cost: $g_{new}(s_j) = g(s_i) + \sqrt{\Delta s^T \mathcal{M}^*(s_i, t) \Delta s}$.
- **Terminate:** When s_{goal} is reached. The traced path corresponds to the optimal discrete geodesic on the semantically-warped manifold.

This precise discretization strategy ensures that the discrete graph search strictly respects the continuous, anisotropic constraints of the physical world, maintaining the $O(N \log N)$ efficiency required for macro-scale deployment.

E Extended Implementation Source Code

E.1 Automated Causal Knowledge Distillation Prompts

To extract structured ecological priors from unstructured natural language text, we utilize a schema-guided few-shot prompting strategy. The Large Language Model (e.g., Gemini 2.5 Pro) acts as a domain expert, distilling descriptive biology into mathematically tractable potential field parameters (α for amplitude, γ for spatial decay slope). This process relies on structured generation APIs to guarantee deterministic JSON outputs.

```
SYSTEM_PROMPT = """
You are an expert computational ecologist. Your task is to extract spatial and
behavioral constraints from unstructured biological text and convert them into
deterministic mathematical potential parameters for a Riemannian planner.

For each semantic feature, you must strictly extract:
 - Target Habitat: The discrete environmental class (e.g., Water, Marsh, Road).
 - Behavioral Type: Whether the feature causes 'Attraction' or 'Repulsion'.
 - Amplitude (alpha): The strength of the attraction/repulsion [0.0, 1.0].
 - Decay Slope (gamma): The spatial decay rate over distance [0.01, 1.0].
 - Weather Sensitivity: How vulnerable the species is to extreme weather [0.0, 1.0].

Respond strictly using the provided JSON schema.
"""

FEW_SHOT_EXAMPLE = """
User: "The Black-necked Crane strongly prefers foraging in shallow marshes but
is extremely terrified of paved human roads, fleeing if within 50 meters."

Assistant:
{
  "species_name": "Black-necked Crane",
  "habitat_potentials": [
    {
      "habitat_class": "Marsh",
      "type": "Attraction",
      "amplitude_alpha": 0.95,
      "decay_slope_gamma": 0.10
    },
    {
      "habitat_class": "Road",
      "type": "Repulsion",
      "amplitude_alpha": 1.0,
      "decay_slope_gamma": 0.50
    }
  ],
  "weather_sensitivity": 0.8
}
"""

USER_PROMPT = """
Extract the spatial parameters for the following field observation:
"{field_text_input}"
"""
```

E.2 Neuro-Symbolic JSON Schema

To bridge the gap between stochastic LLM text generation and rigorous geometric modulation, the outputs are constrained via `Pydantic` models. This schema serves as a gatekeeper, ensuring that the extracted semantic values remain bounded within valid ranges, thus preventing unbounded tensor deformations (e.g., infinite spatial scaling) before the optimization phase.

```
from pydantic import BaseModel, Field
from typing import List, Literal

class SpatialPotential(BaseModel):
    habitat_class: Literal["Forest", "Water", "Marsh", "Road", "Dirt", "Shrub"]
    type: Literal["Attraction", "Repulsion"]

    amplitude_alpha: float = Field(
        ..., ge=0.0, le=1.0,
        description="Maximum intensity of the potential field well or barrier."
    )
    decay_slope_gamma: float = Field(
        ..., ge=0.01, le=1.0,
        description="Exponential decay rate controlling the gradient smoothness."
    )

class BirdEcologicalPrior(BaseModel):
    species_name: str
    habitat_potentials: List[SpatialPotential]

    weather_sensitivity: float = Field(
        ..., ge=0.0, le=1.0,
        description="Scaling factor applied to potentials during adverse weather."
    )
```

E.3 Core Geodesic Solver and Conformal Scaling Implementation

The core of the Causal-Geo framework relies on conformally scaling the base physical metric tensor $\mathbf{M}$ using the generated semantic potential field U_{eco}. Below, we present the core computational logic for the anisotropic edge cost calculation ($\sqrt{\Delta s^T \mathcal{M}^* \Delta s}$) used to construct the sparse graph for our discrete geodesic A^* solver. Note that in actual deployment, this loop is accelerated via Numba `@njit` to handle the macro-scale 12.2 km^2 grids efficiently.

```
import numpy as np
from scipy.sparse import csr_matrix
from numba import njit

# Note: @njit used in production for C-level speed
@njit
def compute_conformal_edges(h, w, M_xx, M_yy, M_xy, U_eco_matrix):
    """
    Dynamically stretches spatial distance based on the Semantic Potential Field.
    Evaluates the discrete Riemannian quadratic form: ds^T * M * ds
    """
    max_edges = h * w * 8
    row, col, data = np.zeros(max_edges), np.zeros(max_edges), np.zeros(max_edges)

    # 8-connected grid directions: (dy, dx)
    directions = [(0, 1), (1, 0), (0, -1), (-1, 0),
                  (1, 1), (1, -1), (-1, 1), (-1, -1)]

    idx = 0
    for y in range(h):
        for x in range(w):
            u = y * w + x
            u_eco_u = U_eco_matrix[y, x]

            # Obstacle avoidance early stopping (infinite metric constraint)
            if M_xx[y, x] > 1e5 or M_yy[y, x] > 1e5:
                continue

            for dy, dx in directions:
                ny, nx = y + dy, x + dx
                if 0 <= nx < w and 0 <= ny < h:
                    u_eco_v = U_eco_matrix[ny, nx]

                    if M_xx[ny, nx] > 1e5 or M_yy[ny, nx] > 1e5:
                        continue
```

```
                    v = ny * w + nx

                    # 1. Average the metric tensor along the edge
                    m_xx = (M_xx[y, x] + M_xx[ny, nx]) / 2.0
                    m_yy = (M_yy[y, x] + M_yy[ny, nx]) / 2.0
                    m_xy = (M_xy[y, x] + M_xy[ny, nx]) / 2.0

                    # 2. Evaluate Quadratic Form: ds^T * M * ds
                    ds_sq = m_xx*(dx**2) + 2.0*m_xy*(dx*dy) + m_yy*(dy**2)

                    # Riemannian approximation: base physical metric
                    base_physics_cost = np.sqrt(max(ds_sq, 1e-8))

                    # 3. Conformal Factor calculation (U_eco)
                    conformal_factor = (u_eco_u + u_eco_v) / 2.0

                    # 4. Warped Riemannian Edge Cost: U_eco * sqrt(ds^T M ds)
                    edge_cost = conformal_factor * base_physics_cost

                    row[idx] = u
                    col[idx] = v
                    data[idx] = edge_cost
                    idx += 1

    return row[:idx], col[:idx], data[:idx]
```

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction explicitly state the framework's capabilities and appropriately bound our claims to the evaluated macroscopic and microscopic environments.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: We explicitly discussed the limitations of our approach, including reliance on perception fidelity and discretization bounds, in the concluding section.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The assumption and statement of Theorem 1 (Completeness and Optimality) are provided in Section 3.3, with the complete formal proof detailed in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We detail the experimental setup, baseline configurations, and discretization parameters in the main text and appendices to ensure reproducibility.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We provide the 150 test queries used for evaluation in the supplementary material. As discussed in the Broader Impacts section, due to potential dual-use risks associated with autonomous physical navigation, the core execution engine is withheld to ensure safe deployment and prevent unauthorized usage.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.

- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Implementation details, including the specific LLM prompting schema and mathematical mapping parameters, are thoroughly documented in Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report the mean and standard deviation for our quantitative metrics (OUS, LVR, Path Cost) across 150 stratified trials in Table 1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

Justification: The inference time and scalability metrics are reported in Table 3, and the hardware specifications are detailed in the supplementary material.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: Our research fully conforms to the NeurIPS Code of Ethics, focusing on safe, transparent, and physically grounded autonomous planning.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We briefly discuss the broader impacts, particularly regarding how our physical gatekeeper prevents autonomous hazards, alongside the limitations in the conclusion.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper proposes a geometric planning framework and does not release high-risk assets like pre-trained generative foundation models or scraped datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The orthophoto dataset used for the wetland simulator is properly cited in the references, and its terms of use are respected.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The macro-scale digital twin environment and the corresponding evaluation scenarios introduced in this work are documented in the supplementary material.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This research focuses on embodied AI planning algorithms and does not involve crowdsourcing or human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This research does not involve human subjects, hence IRB approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The usage of Gemini 2.5 Pro as the core semantic reasoning engine for Automated Causal Knowledge Distillation is explicitly detailed in Section 4.2 and Appendix B.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.

- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.