drones

MDPI

Article

A Fine-Grained Difficulty and Similarity Framework for Dynamic Evaluation of Path-Planning Generalization in UGVs

Zewei Dong [1,2,*], Yaze Guo [2], Jingxuan Yang [1], Xiaochuan Tang [2], Weichao Xu [1] and Ming Lei [1]

[1] Department of Automation, Tsinghua University, Beijing 100084, China; yangjx20@mails.tsinghua.edu.cn (J.Y.)
[2] Department of Avionics and Ordnance Engineering, Army Aviation Institute, Beijing 100123, China
* Correspondence: dzw22@mails.tsinghua.edu.cn

Highlights

What are the main findings?

- This study proposes a novel dual-axis evaluation framework that deconstructs UGV generalization into quantifiable dimensions of scenario similarity and intrinsic difficulty. It provides a diagnostic lens to separate failures due to lack of robustness from those due to generalization limits.
- The framework is powered by two methodological pillars: a hierarchical four-layer similarity quantification method and a novel hybrid consensus mechanism for objective difficulty annotation, enabling the construction of an interpretable 3D performance landscape for diagnostic evaluation.

What are the implications of the main finding?

- The framework provides a diagnostic tool that identifies the root cause of model failure by distinguishing between insufficient robustness and fundamental generalization limits, directly supporting safer UGV deployment and informed model/algorithm selection and safety certification.
- It establishes a standardized, dynamic evaluation paradigm. The interpretable performance landscape and metrics enable root-cause analysis for model improvement and informed, safety-aware deployment decisions in unseen environments.

Abstract

The generalization capability of the decision-making modules in unmanned ground vehicles (UGVs) is critical for their safe deployment in unseen environments. Prevailing evaluation methods, which rely on aggregated performance over static benchmark sets, lack the granularity to diagnose the root causes of model failure, as they often conflate the distinct influences of scenario similarity and intrinsic difficulty. To overcome this limitation, we introduce a fine-grained, dynamic evaluation framework that deconstructs generalization along the dual axes of multi-level difficulty and similarity. First, scenario similarity is quantified through a four-layer hierarchical decomposition, with results aggregated into a composite similarity score. Test scenarios are independently classified into ten discrete difficulty levels via a consensus mechanism integrating large language models and task-specific proxy models. By constructing a three-dimensional (3D) performance landscape across similarity, difficulty, and task performance, we enable detailed behavioral diagnosis. The framework assesses robustness by analyzing performance within the high-similarity band (90–100%), while the full 3D landscape characterizes generalization under distribution shift. Seven interpretable metrics are derived to quantify distinct facets of both generalization and robustness. This initial validation focuses on the path-planning layer under

Academic Editor: Yushu Yu

Received: 23 December 2025
Revised: 22 January 2026
Accepted: 27 January 2026
Published: 31 January 2026

full state observability, establishing a foundational proof-of-concept for the framework. It not only ranks algorithms but also reveals non-trivial behavioral patterns, such as the decoupling between in-distribution robustness and out-of-distribution generalization. It provides a reliable and interpretable foundation for evaluating the readiness of UGVs for safe deployment in unseen environments.

Keywords: generalization capability; unmanned ground vehicles (UGVs); dynamic evaluation; robustness

1. Introduction

The intelligent decision-making systems of unmanned ground vehicles (UGVs) rely on learning from extensive scenario data [1–4]. However, the safety and reliability of these systems in real-world deployment fundamentally depend on their capability to transfer and adapt learned knowledge to unseen scenarios—that is, their generalization capability. A rigorous, fine-grained evaluation of this capability is crucial. This evaluation directly determines the operational boundaries and safety limits of UGV intelligence [5,6].

Current research on evaluating autonomous systems has progressed along several key directions [7]. Prominent benchmarks and simulation frameworks [8,9] have established standardized test suites, providing essential baselines for comparing different algorithms. A growing focus on out-of-distribution (OOD) generalization has led to studies highlighting its critical role [10,11] and to the creation of specialized datasets designed to probe specific failure modes [12]. In parallel, the robustness of models has been investigated, often through the lens of performance under adversarial perturbations or sensor noise [13]. To better characterize test conditions, methods have been proposed to quantify scenario complexity or difficulty, employing both rule-based metrics [14] and learning-based estimators [15]. Furthermore, concepts of dynamic evaluation and continual testing are emerging to assess model adaptation over time or across sequential scenarios [15–17]. Specific to ground vehicles, the most recent works not only propose metrics for online reliability estimation [18] and generate challenging scenarios for robustness stress-testing [19] but also employ time series analysis for dynamic assessment of learning internalization capability [20]. Despite these valuable efforts, the evaluation landscape remains fragmented; a cohesive framework that jointly and explicitly quantifies the multi-faceted nature of scenario similarity and intrinsic difficulty to provide a fine-grained, interpretable diagnosis is still missing [21–25], particularly for UGV-specific deployment safety [26].

In both research and practice, the prevailing evaluation paradigm still predominantly relies on static benchmark sets [27]. Aggregate performance metrics (e.g., average success rate) computed over a fixed collection of test scenarios are used to infer generalization. This paradigm, while practical, suffers from fundamental limitations: it fails to disentangle whether performance degradation arises from true distribution shift (a generalization issue) or from instability under minor variations within a similar distribution (a robustness issue); it does not systematically measure or control the similarity gradient between test and training data, obscuring the continuous effect of distribution shift; and it overlooks the independent role of scenario difficulty, conflating performance changes due to increased complexity with those caused by distributional mismatch [28,29]. The nuanced relationship between these two capabilities has been a topic of recent discussion in the robustness-generalization literature for deep learning systems [30].

Despite these valuable efforts, a unified framework is still missing. This framework would need to measure two things at once: how similar a test scenario is to the training data, and how difficult that scenario is on its own. Because this framework is absent,

current evaluation methods have clear limitations. For example, static benchmark sets look only at the final average score. This mixes together two different problems: whether the data distribution has changed, and how complex the scenario is. Other methods measure only the difficulty of a scenario but ignore how similar it is to what the model was trained on. This makes it hard to tell apart a very hard example from the training distribution and a truly new, unseen type of example. Similarly, specialized OOD benchmarks test how models handle new types of data. However, they often do not carefully check how robust a model is to small changes within familiar data. They also usually do not separate the effects of difficulty and similarity clearly. This situation makes it hard to find the real cause of a model's failure. Does it fail because it is not robust to small changes in familiar situations? Or does it fail because it has reached its limit in handling something completely new? Our framework is designed to solve this exact problem. We separate generalization into two independent dimensions: similarity and difficulty. This allows for a clearer diagnostic evaluation that current methods cannot provide. A conceptual comparison of existing evaluation methods is presented in Table 1.

Table 1. Conceptual comparison of evaluation methods for UGV generalization.

Evaluation Paradigm	Key Focus	Decouples Similarity & Difficulty?	Diagnostic Granularity	Dynamic Assessment?
Static Benchmarks	Aggregate performance over a fixed set	No	Low	No
Difficulty-Only Metrics	Intrinsic scenario complexity	No	Medium	Potentially
OOD-Focused Benchmarks	Performance under distribution shift	Often No	Medium	Limited
Proposed Framework	Dual-axis	Yes	High	Yes

Therefore, we introduce a novel dynamic evaluation paradigm. This paradigm moves beyond aggregated metrics via a fine-grained decomposition along the axes of difficulty and similarity. To demonstrate and validate this general framework, we instantiate it for the fundamental task of UGV path planning and examine its diagnostic power in a comparative study of representative algorithms. This approach provides a concrete experimental validation while maintaining the framework's conceptual generality. The core innovations of our framework are:

(1) Multi-dimensional Similarity Quantification: Scenarios are deconstructed into four hierarchical layers—natural environment, static objects, dynamic objects, and semantic information—for comprehensive similarity assessment, capturing both low-level physical features and high-level rule-based characteristics.
(2) Multi-level Difficulty Annotation: A consensus mechanism integrating large language models and task-specific proxy models is introduced to independently classify test scenarios into ten discrete difficulty levels, enabling finer characterization of intrinsic scenario complexity.
(3) Dynamic three-dimensional(3D) Evaluation: A 3D performance landscape across similarity intervals, difficulty levels, and task performance is constructed to enable in-depth behavioral diagnosis. Specifically, we propose analyzing performance within the high-similarity band (90–100%) to specifically evaluate system robustness, while the full landscape characterizes generalization under distribution shift.

(4) Comprehensive Performance Metrics: Seven interpretable quantitative metrics are extracted from the 3D landscape to comprehensively measure different facets of a model's generalization and robustness.

To clearly demonstrate and validate the core methodological pillars of our proposed framework, we instantiate it in this paper for the fundamental task of path planning under full observability. This choice allows us to focus on the evaluation methodology itself by controlling the complexity introduced by perceptual noise and partial observability, which are critical yet orthogonal challenges for generalization. The framework's design is, however, general and its layers are structured to incorporate perceptual inputs in future extensions.

The contribution of this work lies in systematically identifying the shortcomings of current static evaluation methods and proposing a complete, operational dynamic evaluation solution. Our framework not only quantifies and compares the generalization and robustness of different algorithms for ranking purposes but also reveals the specific patterns of performance degradation, providing a reliable and interpretable analytical foundation for model improvement and system safety certification.

The paper proceeds as follows. Section 2 formulates the problem. Section 3 details the proposed evaluation framework. Section 4 introduces the experimental setup. Section 5 presents the results. Section 6 provides a discussion of the findings, and Section 7 concludes with future work.

2. Problem Description

This paper adopts a systemic definition for the generalization capability of an intelligent decision-making system, encompassing a continuous spectrum from in-distribution stability to out-of-distribution adaptation [7,23,28]. This spectrum is operationalized through two complementary analytical dimensions:

- In-Distribution (ID) Robustness refers to a model's ability to preserve stable performance when faced with scenarios that are highly similar to the training distribution but contain minor perturbations or higher intrinsic complexity. It defines the lower bound and stability foundation of generalization, reflecting the system's resistance to variations within the core characteristics of the training domain.
- Out-of-Distribution (OOD) Generalization refers to a model's capacity to maintain its decision-making performance when encountering novel scenarios that exhibit a significant, core shift from the training data distribution. It defines the boundary and extensibility of generalization, reflecting the system's potential to transfer acquired knowledge to fundamentally new situations.

Our dynamic evaluation framework explicitly integrates ID robustness analysis as a key diagnostic component within the broader generalization assessment. By establishing a high-similarity band for dedicated robustness evaluation, we can clearly distinguish performance decay due to model instability from that arising from encountering fundamentally novel scenarios. This integrated diagnostic perspective is a core advantage of our framework over isolated evaluation approaches.

The following sections formalize the core concepts of this framework in general terms applicable to various UGV decision-making tasks. For concreteness, we illustrate and evaluate these concepts using a path-planning task, where the performance metric $P(M, s)$ is instantiated as the success rate of navigating from a start to a goal point. This choice does not limit the framework's applicability to other tasks.

A rigorous evaluation framework must disentangle the two facets of generalization outlined above. The core limitation of conventional evaluation lies in its reliance on a single aggregate metric over a static test set D_{test}:

$$\bar{P} = \mathbb{E}_{s \sim D_{\text{test}}}[P(M, s)]. \tag{1}$$

where $P(M, s)$ is the task performance of model M on scenario s. This monolithic approach conflates the distinct influences of scenario similarity and intrinsic difficulty.

2.1. Scenario Similarity

An operational scenario s for a UGV decision-making task is structured into four hierarchical layers:

$$s = (s_{\text{env}}, s_{\text{sta}}, s_{\text{dyn}}, s_{\text{sem}}). \tag{2}$$

representing natural environment, static objects, dynamic objects, and semantic information, respectively. The overall similarity $\Phi(s)$ of a test scenario to the training distribution D_{train} is defined as its maximum weighted similarity across all layers to any scenario in the training set:

$$\Phi(s) = \max_{s' \in D_{\text{train}}} \sum_{l \in L} w_l \cdot \phi_l(s, s'). \tag{3}$$

where ϕ_l is the layer-specific similarity metric, $L = \{\text{env, sta, dyn, sem}\}$, and $\sum w_l = 1$. This multi-layer quantification captures both low-level physical and high-level semantic features.

We define distribution shift as a continuous decrease in scenario similarity. Our test set spans this spectrum. Extensive perturbations to geometric and dynamic parameters create low-similarity scenarios, presenting a quantifiable out-of-distribution challenge. This allows us to study gradual performance decay with increasing distribution shift. Generalization to topologically distinct environments represents a broader OOD class; our framework's layers can capture such differences, and their evaluation is planned for future work.

2.2. Scenario Difficulty

We introduce an independent dimension, scenario difficulty $\Psi(s) \in \{1, \ldots, 10\}$, which quantifies the inherent complexity of the decision logic required in s, irrespective of its distributional properties. Critically, we assume $\Psi(s)$ is approximately statistically independent of $\Phi(s)$, i.e., $\rho(\Phi, \Psi) \approx 0$. This means a highly similar scenario can be very difficult, and a very dissimilar scenario can be relatively simple.

This independence is a reasonable and useful assumption for diagnostic evaluation. The similarity $\Phi(s)$ quantifies content-based alignment with the training data, reflecting what elements are present. The difficulty $\Psi(s)$ quantifies the complexity of the required decision logic, reflecting how hard it is to succeed given those elements. These two dimensions are conceptually distinct. For instance, a scenario densely populated with obstacles seen during training (high Φ) can be extremely challenging to navigate (high Ψ) due to the intricate layout. Conversely, a scenario containing novel object types (low Φ) might offer ample free space and simple dynamics, making it easy (low Ψ). Treating them as approximately independent axes allows our framework to separate the influence of distribution shift from that of intrinsic task complexity, enabling precise root-cause analysis.

2.3. High-Similarity Band for Robustness Assessment

To explicitly decouple ID Robustness from OOD Generalization, we define a high-similarity band using a conservative threshold $\gamma = 0.9$:

$$S_{\text{high-sim}} = \{s \in D_{\text{test}} \mid \Phi(s) \geq \gamma\}. \quad (4)$$

The choice of $\gamma = 0.9$ is grounded in methodological rigor and practical convention. This high threshold ensures that scenarios within $S_{\text{high-sim}}$ are, with high confidence, aligned with the core characteristics of the training distribution. Therefore, significant performance degradation within this band can be more robustly attributed to a model's sensitivity to minor in-distribution perturbations (i.e., a lack of robustness) rather than to a fundamental distribution shift [7]. This operational definition aligns with established practice in robotic reliability assessment [31,32] and treats γ as a configurable safety parameter. The choice of a high threshold (e.g., 0.9 or 0.95) to isolate in-distribution performance is also supported by recent methodologies for safety-critical autonomous system validation [33]. Our experiments confirm that the core conclusions remain stable under minor variations of γ.

Consequently, a model's performance profile within $S_{\text{high-sim}}$ specifically evaluates its in-distribution robustness, while its performance across the full Φ spectrum characterizes out-of-distribution generalization.

2.4. Core Evaluation Objectives

Formally, our framework is designed to address the following interconnected objectives:

(1) Similarity Spectrum Partitioning: Systematically partition the test set D_{test} into intervals based on the values of $\Phi(s)$.
(2) Independent Difficulty Annotation: Objectively assign a discrete difficulty level $\Psi(s))$ to each test scenario using an automated, objective process independent of human judgment.
(3) 3D performance landscape Construction: Construct a diagnostic performance landscape $P = f(\Phi, \Psi)$ from the set of tuples $\{(\Phi(s), \Psi(s), P(M, s))\}$.
(4) Decoupled Metric Extraction: Derive a set of interpretable metrics from this surface to separately quantify: ID Robustness and OOD Generalization.

This decoupling allows independent assessment of both robustness and generalization. Furthermore, it establishes a foundation for us to probe complex relationships between these capabilities. These relationships may include trade-offs or conflicts.

The following sections present our methodology to achieve these objectives, providing a fine-grained, diagnostic tool for evaluating UGV generalization.

3. Methodology

This section details the proposed fine-grained dynamic evaluation framework for generalization capability. The core philosophy of our method is to transcend the traditional evaluation paradigm that relies on a single aggregated metric (e.g., average test performance). By systematically deconstructing test scenarios and performing multi-level quantification along two orthogonal dimensions—similarity and difficulty—we achieve a leap from coarse-grained ranking to fine-grained root-cause diagnosis. As illustrated in Figure 1, the framework consists of three sequentially executed and logically rigorous stages: Scenario Similarity Quantification, Scenario Difficulty Classification, and 3D Dynamic Evaluation.

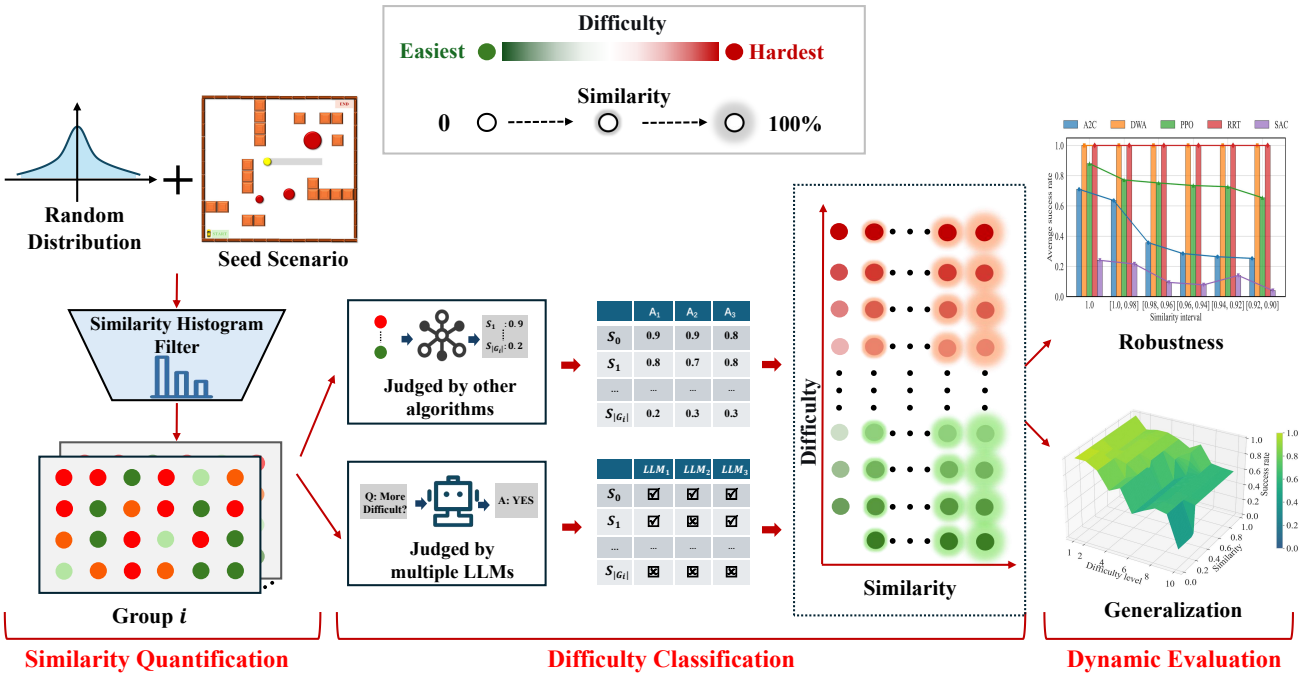

Figure 1. The proposed dual-axis evaluation framework deconstructs generalization into similarity and difficulty to construct a diagnostic 3D performance landscape.

This study constructs a systematic evaluation framework, as shown in Figure 1.

3.1. Similarity Quantification

This stage aims to compute a continuous similarity score $\Phi(s) \in [0,1]$ for each test scenario s relative to the training distribution D_{train}. To achieve a comprehensive and structured measurement of scenario content, we propose a hierarchical deconstruction and matching method across four distinct layers.

3.1.1. Hierarchical Feature Extraction and Similarity Computation

Each scenario s is parsed into a quadruple: $s = (s_{\text{env}}, s_{\text{sta}}, s_{\text{dyn}}, s_{\text{sem}})$. For each layer, a specific similarity metric function $\phi_l(\cdot,\cdot)$ is defined to compute the matching degree between a test scenario and an individual training scenario s' at that layer.

Natural Environment Layer: This layer describes physical environmental conditions, formalized as a discrete state vector, e.g., $s_{\text{env}} = (W, L, R)$, representing weather, illumination, and road surface conditions, respectively. For each dimension $d \in \{W, L, R\}$, a normalized similarity is calculated based on a predefined distance matrix $D_{\text{env}}^{(d)}$:

$$\phi_{\text{env}}^{(d)}(s, s') = 1 - \frac{D_{\text{env}}^{(d)}(s_d, s'_d)}{\max(D_{\text{env}}^{(d)})}. \tag{5}$$

The comprehensive similarity for this layer is obtained by arithmetic averaging:

$$\phi_{\text{env}}(s, s') = \frac{1}{3} \sum_{d \in \{W,L,R\}} \phi_{\text{env}}^{(d)}(s, s'). \tag{6}$$

Static Object Layer: This layer consists of a set of static objects, $s_{\text{sta}} = \{o_i \mid i = 1, \ldots, Q\}$, where each object o_i contains attributes such as shape, position, and size. We compute

the similarity between static object sets via optimal global matching. First, we define a matching score for any object pair o_i and o'_j :

$$\phi_{\text{obj}}(o_i, o'_j) = I_{\text{shape}}(o_i, o'_j) \cdot \exp\left(-\frac{\|\text{pos}_i - \text{pos}'_j\|^2}{2\sigma^2_{\text{pos}}}\right), \tag{7}$$

Here, I_{shape} indicates shape compatibility. Next, we apply the Hungarian algorithm to the resulting similarity matrix. This finds the maximum-weight matching M^*. Finally, we incorporate a penalty for differences in set size. The overall layer similarity is:

$$\phi_{\text{sta}}(s, s') = \frac{\sum_{(i,j)\in M^*} \phi_{\text{obj}}(o_i, o'_j)}{\max(Q, Q')}. \tag{8}$$

Dynamic Object Layer: This layer describes the behavior of moving objects, $s_{\text{dyn}} = \{m_j \mid j = 1,\ldots,N\}$. Each dynamic object m_j is characterized by its trajectory $traj_j$, velocity v_j, and direction dir_j. We fuse two components to calculate similarity for a pair of dynamic objects. The first component is trajectory similarity. We measure it using the F1-score of the Continuous Longest Common Subsequence (LCS). Its formula is:

$$\phi_{\text{traj}}(traj_j, traj'_k) = \frac{2 \cdot |\text{LCS}_{\text{cont}}(traj_j, traj'_k)|}{|traj_j| + |traj'_k|}. \tag{9}$$

The second component is motion vector similarity. This assumes uniform linear motion. It is calculated as:

$$\phi_{\text{motion}}(m_j, m'_k) = \exp\left(-\frac{|v_j - v'_k|}{\sigma_v} - \frac{\angle(dir_j, dir'_k)}{\sigma_{dir}}\right). \tag{10}$$

The overall similarity $\phi_{\text{dyn}}(s, s')$ is then computed using a global matching strategy analogous to the static layer.

Semantic Information Layer: This layer is designed within the framework to capture high-level rules and functional context, which are crucial for generalization evaluation in complex, semantically-rich environments [34]. In general, the semantic layer can be formalized as a k-dimensional feature vector $s_{\text{sem}} = (a_1, a_2,\ldots,a_k)$, where $a_m \in \{0,1\}$ indicates the activation of the m-th semantic rule (e.g., "must yield to pedestrians", "speed limit zone").

Implementation in This Study:Due to the current lack of a perception module in our experimental setup (see Section 4), we focus on a subset of semantic rules that can be derived directly from the ground-truth state information. Specifically, we define $k = 2$ rules relevant to our path-planning task: "narrow passage exists" and "dynamic trajectory conflict", as detailed in Table 2. The similarity between the two scenarios for this layer is computed using a weighted Jaccard coefficient:

$$\phi_{\text{sem}}(A, B) = \frac{\sum_{m=1}^{k} \omega_m \cdot \min(a_m, b_m)}{\sum_{m=1}^{k} \omega_m \cdot \max(a_m, b_m)}. \tag{11}$$

where ω_m is a weight reflecting the relative importance of the m-th rule. Given that both selected rules represent significant navigation challenges (spatial constraint and interactive conflict), and to avoid introducing unnecessary bias in this proof-of-concept study, we adopt an equal weighting scheme: $\omega_1 = \omega_2 = 0.5$. This design keeps the semantic layer functional and interpretable within the current simulation constraints, while the

framework's architecture readily accommodates the addition of more complex, perception-dependent rules in future work.

Table 2. Four-layer hierarchical feature decomposition of the benchmark seed scenario.

Feature Layer	Description Item	Specific Parameters and Notes
Natural Environment	Weather and Illumination (Fixed)	Weather: Clear; Illumination: Daytime; Road Surface: Dry. For simplified analysis, the environmental layer is fixed in this experiment. This layer remains constant across all scenarios.
Static Objects	Obstacle 1 Obstacle 2 Obstacle 3	Type: Spherical obstacle; Position: $[9.0, 6.5]$; Radius: 1.0 m. Type: Spherical obstacle; Position: $[13.0, 7.5]$; Radius: 1.5 m. Type: Spherical obstacle; Position: $[17.0, 16.0]$; Radius: 2.0 m.
Dynamic Objects	Moving Obstacle	Type: Dynamic spherical obstacle; Radius: 1.0 m; Motion pattern: Uniform back-and-forth along the x-axis; Trajectory endpoints: $[10, 12]$ and $[18, 12]$; Velocity: 1.0 m/s.
Semantic Information	Rule 1: Narrow Passage Existence Rule 2: Dynamic Trajectory Conflict	Condition: Present if the distance between any two static obstacles is less than a predefined safety threshold (e.g., 2 × UGV_width). In the seed scenario: Obstacle_1 and Obstacle_2 form a narrow passage. Condition: Present if the predicted trajectory of the dynamic obstacle intersects the straight-line path from start to goal. In the seed scenario: The dynamic obstacle's path between $[10, 12]$ and $[18, 12]$ intersects the ideal UGV path.
Task Ontology	UGV Parameters Start and Goal Points	Dimensions: 1.0 m × 0.5 m; Drive mode: Differential drive; Maximum steering angle: 30°. Start: $[1, 1]$; Goal: $[23, 23]$.

In this initial study, the semantic layer incorporates a minimal but representative set of rules directly derivable from ground-truth state to demonstrate its integration into the similarity framework. The primary purpose here is to establish the layer's role as a structured, extensible component within the hierarchical model. The framework is designed to readily accommodate more complex, perception-driven semantic features for evaluating generalization in richer environments, which is a key direction for future work.

3.1.2. Overall Similarity Fusion

The final similarity score $\Phi(s)$ of a given test scenario s relative to the entire training set is obtained by identifying its "most similar" counterpart s' within D_{train} and fusing the layer-wise results:

$$\Phi(s) = \max_{s' \in D_{\text{train}}} \left(\sum_{l \in L} w_l \cdot \phi_l(s, s') \right). \tag{12}$$

Here, $L = \{\text{env}, \text{sta}, \text{dyn}, \text{sem}\}$, and w_l are configurable weights for the respective layers, satisfying $\sum w_l = 1$. These weights can be set using methods like the Analytic Hierarchy Process (AHP) according to specific task requirements. This step assigns a continuous and interpretable similarity value to each test scenario, forming the crucial "distribution shift gradient" for subsequent analysis. Having defined similarity quantification, we now introduce the independent axis of scenario difficulty.

The layer weights w_l in Equation (12) are configurable to reflect task-specific priorities [35]. For this study, we adopted a balanced weighting scheme ($w_{\text{env}} = 0.2$, $w_{\text{sta}} = 0.3$, $w_{\text{dyn}} = 0.3$, $w_{\text{sem}} = 0.2$) after a sensitivity analysis conducted on a held-out validation set of scenarios. This scheme slightly emphasizes object-level and rule-level features, which were found to be more discriminative for the path-planning task than environmental conditions in our setup. The sensitivity analysis confirmed that the relative ranking of algorithms was robust to moderate variations (± 0.1) in these weights.

In summary, the hierarchical similarity quantification provides a structured way to measure distributional shift. For the UGV path-planning task here, the most important factors are the static and dynamic object layers. These layers capture obstacle layout and motion. Task-relevant semantic rules, such as those for narrow passages, are also key. This is shown by their assigned weights. The proposed framework is general. Its core ideas are the four-layer decomposition, the maximum-similarity fusion method, and separating similarity from difficulty. The specific parts can be customized for different tasks. These include the similarity metrics for each layer, the layer weights, the definitions of semantic rules, and the jury for difficulty annotation. This design lets the framework adapt to other UGV decision-making tasks. Users can change these specific elements while keeping the main diagnostic structure.

3.2. Difficulty Classification

The flow diagram is illustrated in Figure 2. To independently and objectively quantify the decision complexity of a scenario, this framework proposes a hybrid consensus mechanism that synergistically integrates two complementary juries: a LLM-based jury [36,37] providing a priori reasoning estimates and a proxy-model-based jury [38–40] deriving a posteriori performance evidence. This dual-path design ensures the objectivity, reproducibility, and fine-grained discriminative power of the resulting annotations. Difficulty is divided into 10 discrete levels (Level 1 being the simplest, Level 10 the most difficult), thereby enabling a structured and nuanced characterization of the "scenario difficulty" dimension.

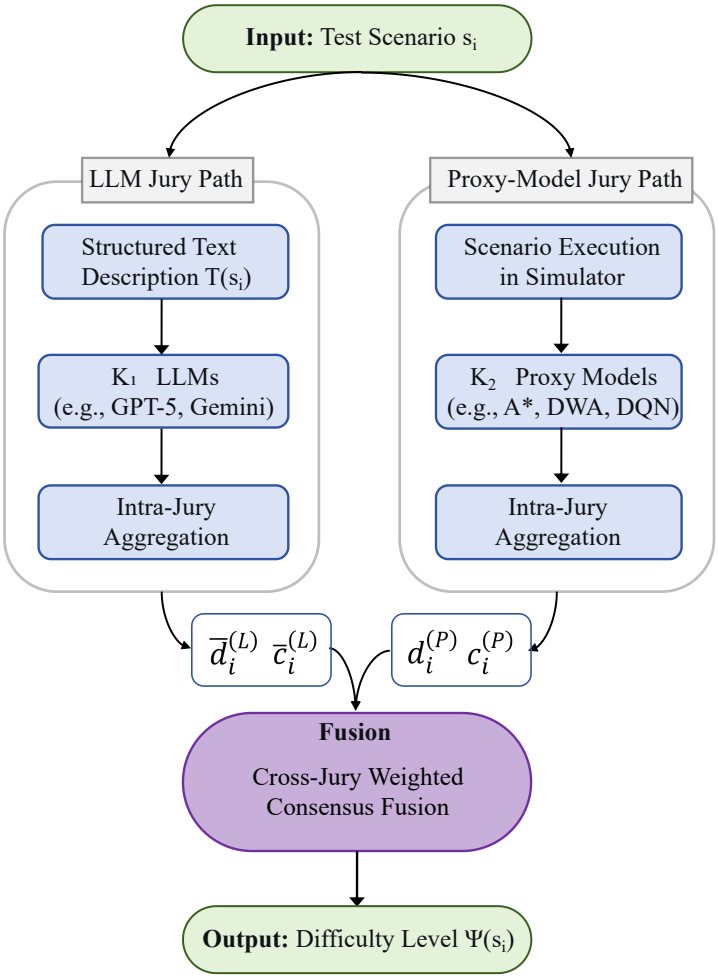

Figure 2. Flow diagram of the hybrid consensus mechanism for objective difficulty annotation.

3.2.1. Multi-Jury Consensus Mechanism

(1) Structured Scenario Description.
The four-layer features of each test scenario s_i are automatically encoded into a standardized structured description $T(s_i)$, which serves as the input for the LLM jury. For the proxy model jury, the scenario s_i is directly executed in the simulation environment based on its underlying parameters.

(2) Dual-Path Jury Composition.
LLM Jury (A Priori Reasoning Difficulty): Comprises K_1 Large Language Models (e.g., GPT-5, DeepSeek-V3.2). Each LLM receives a structured prompt. This prompt contains the scenario description $T(s_i)$ and a standardized instruction. The instruction asks the LLM to assess the navigation difficulty on a scale from 1 to 10. The assessment should consider several key factors. These factors include obstacle density, spatial constraints, dynamic interaction complexity, and rule compliance. The prompt for the LLM jury-based difficulty assessment is shown in Figure 3.

LLM Jury Prompt for Difficulty Assessment

Role: You are an expert in autonomous vehicle navigation and scenario analysis, specializing in evaluating the decision-making complexity for unmanned ground vehicles (UGVs). Your task is to assess the intrinsic difficulty of a given navigation scenario based on its structured description.

Scenario Description Input Format:
- **Natural Environment Layer:** weather, illumination, road surface.
- **Static Object Layer:** type, position, size of static obstacles.
- **Dynamic Object Layer:** trajectory, velocity, direction of moving obstacles.
- **Semantic Information Layer:** high-level rules (e.g., narrow passage, trajectory conflict).

Your Task:
Analyze the provided scenario and estimate its navigation difficulty on a discrete scale from 1 (easiest) to 10 (hardest). Consider the following factors in your assessment:
-**Obstacle density and layout complexity**
-**Spatial constraints** (e.g., narrow passages)
-**Dynamic interaction complexity** (e.g., moving obstacles, trajectory conflicts)
-**Rule compliance requirements** (e.g., yielding, speed limits)
-**Overall decision-making uncertainty for a UGV agent**

Output Format Requirements:
You must return your assessment in the following exact JSON structure:

```
json
{
 "difficulty_level": <integer between 1 and 10>,
 "confidence": <float between 0.0 and 1.0>,
 "reasoning_summary": "<brief explanation of key difficulty factors>"
}
```

#Example Input Scenario (structured text):
- **Natural Environment Layer:** Clear, Daytime, Dry Road.
- **Static Object Layer:** Sphere at [9.0,6.5], radius=1.0m; Sphere at [13.0,7.5], radius=1.5m.
- **Dynamic Object Layer:** Sphere moving between [10,12] and [18,12] at 1.0 m/s.
- **Semantic Information Layer:** Narrow passage present; Dynamic trajectory conflict with UGV path.

Figure 3. Structured prompting for a priori difficulty assessment by the LLM jury.

Each LLM then outputs a difficulty level $d_{ij}^{(L)} \in \{1,\ldots,10\}$ and a confidence score $c_{ij}^{(L)} \in [0,1]$ based on its analysis of $T(s_i)$. We parse the JSON response to extract these values.

Proxy Model Jury (A Posteriori Performance Difficulty): Comprises K_2 benchmark path-planning algorithms, including the algorithms under evaluation, among others. Each algorithm independently executes the task in scenario s_i for N runs, and its average success rate p_{kj} is recorded. The individual difficulty score for that algorithm is calculated as $d_{ij}^{(P)} = \text{Round}(1 + 9 \times (1 - p_{kj}))$. The jury's comprehensive difficulty level $d_i^{(P)}$ and confidence $c_i^{(P)}$ are aggregated from the performance of all baseline algorithms.

(3) Intra-Jury Aggregation.

LLM Jury's Comprehensive Difficulty and Confidence:

$$\bar{d}_i^{(L)} = \frac{\sum_{j=1}^{K_1} c_{ij}^{(L)} \cdot \alpha_j^{(L)} \cdot d_{ij}^{(L)}}{\sum_{j=1}^{K_1} c_{ij}^{(L)} \cdot \alpha_j^{(L)}}, \tag{13}$$

$$\bar{c}_i^{(L)} = \frac{1}{K_1} \sum_{j=1}^{K_1} c_{ij}^{(L)}. \tag{14}$$

where $\alpha_j^{(L)}$ is a predefined model reliability coefficient.

Proxy Model Jury's Comprehensive Difficulty and Confidence:

Comprehensive Difficulty: The arithmetic mean of all baseline algorithm difficulty scores is used to reflect the collective opinion center.

$$d_i^{(P)} = \frac{1}{K_2} \sum_{j=1}^{K_2} d_{ij}^{(P)}. \tag{15}$$

Comprehensive Confidence: This is calculated based on the dispersion (standard deviation) of the collective opinions. Smaller dispersion implies higher confidence. To confine the confidence within $[0, 1]$, the standard deviation is mapped to this range. Given that difficulty levels are integers from 1 to 10, the theoretically possible maximum standard deviation is approximately 4.5 (when scores are uniformly distributed at both extremes). Therefore, a normalized confidence formula is:

$$c_i^{(P)} = \max\left(0, 1 - \frac{s_i^{(P)}}{\sigma_{\max}}\right), \tag{16}$$

where,

$$s_i^{(P)} = \sqrt{\frac{1}{K_2 - 1} \sum_{j=1}^{K_2} (d_{ij}^{(P)} - d_i^{(P)})^2} \tag{17}$$

is the sample standard deviation of the scores, and $\sigma_{\max}$ is a normalization constant, which can be set to 4.5 or adjusted based on the actual score distribution. This formula ensures a confidence of 1 when all algorithm scores are perfectly consistent ($s = 0$), and approaches 0 when the disagreement is extreme.

(4) Cross-Jury Weighted Consensus Fusion.

First, calculate the consistency weight for each type of jury's internal opinion. This weight is inversely proportional to the dispersion (variance) of opinions within that jury—smaller variance leads to higher weight.

Calculate the variance of opinions for each jury type:

$$\sigma_i^{2(L)} = \frac{1}{K_1 - 1} \sum_{j=1}^{K_1} (d_{ij}^{(L)} - \bar{d}_i^{(L)})^2, \tag{18}$$

$$\sigma_i^{2(P)} = \frac{1}{K_2 - 1} \sum_{j=1}^{K_2} (d_{ij}^{(P)} - d_i^{(P)})^2. \tag{19}$$

Calculate the consistency weights (normalized using the inverse of variance):

$$w_{\text{consistency}}^{(L)} = \frac{1/(1 + \sigma_i^{2(L)})}{\sum_{X \in \{L,P\}} 1/(1 + \sigma_i^{2(X)})}, \tag{20}$$

$$w_{\text{consistency}}^{(P)} = 1 - w_{\text{consistency}}^{(L)}. \tag{21}$$

We then fuse the opinions from both juries to obtain the final difficulty level. The fusion uses a weighted average formula:

$$\Psi(s_i) = \text{Round}\left(\frac{\bar{c}_i^{(L)} \cdot w_{\text{consistency}}^{(L)} \cdot \bar{d}_i^{(L)} + c_i^{(P)} \cdot w_{\text{consistency}}^{(P)} \cdot d_i^{(P)}}{\bar{c}_i^{(L)} \cdot w_{\text{consistency}}^{(L)} + c_i^{(P)} \cdot w_{\text{consistency}}^{(P)}} \right). \tag{22}$$

3.2.2. Output and Subsequent Processing

Each test scenario ultimately receives an integer difficulty level $\Psi(s_i) \in \{1, 2, \ldots, 10\}$. The entire test set can be partitioned into mutually exclusive subsets based on this level: $S_{\text{test}}^{(k)} = \{s_i \mid \Psi(s_i) = k\}$. Furthermore, within each similarity bin B_i, high- and low-difficulty subsets can be defined (e.g., defining levels 1–4 as low difficulty and levels 7–10 as high difficulty) to provide structured input for the subsequent construction of the 3D performance landscape.

3.3. Dynamic Evaluation

Following the quantification of similarity $\Phi(s)$ and the classification of difficulty $\Psi(s)$ for each test scenario, this final stage synthesizes these dimensions to construct a comprehensive diagnostic performance landscape and extract interpretable metrics for generalization and robustness.

3.3.1. Construction of the Performance Landscape

For a given model M, every test scenario $s \in D_{\text{test}}$ is mapped to a point in a 3D space defined by the coordinates $(\Phi(s), \Psi(s), P(M, s))$, where $P(M, s)$ is the model's performance metric (e.g., success rate). A continuous performance landscape $P = f(\Phi, \Psi)$ is then interpolated from this discrete set of points. This surface visually and quantitatively represents the model's behavioral response across the entire spectrum of distributional shift (similarity) and task complexity (difficulty), forming the core analytical object for subsequent diagnosis.

3.3.2. Robustness and Generalization Analysis

By evaluating performance separately within the high-similarity band and across the full similarity spectrum, our framework not only quantifies ID robustness and OOD generalization in isolation but also positions their interaction and potential trade-offs as a central object of diagnosis.

(1) Robustness Analysis: The model's In-Distribution (ID) Robustness is specifically assessed by analyzing the performance profile within the high-similarity band

$S_{\text{high-sim}}(\Phi \geq 0.9)$. The trend (e.g., rate of decay) and stability (e.g., variance) of P within this restricted region characterize the model's sensitivity to minor perturbations and variations that remain within the core training distribution.

(2) Generalization Analysis: The model's Out-of-Distribution (OOD) Generalization is characterized by the evolution of the performance landscape across the full similarity range ($\Phi \in [0, 1]$). The global pattern of performance degradation as Φ decreases quantifies the model's capability to adapt to increasingly novel scenarios beyond its training experience.

3.3.3. Extraction of Quantifiable Metrics

A suite of seven interpretable metrics is derived from the 3D performance landscape and its specific cross-sections to provide a multi-faceted quantitative profile of the model's capabilities [41].

(1) Attenuation Degree (AD): Quantifies the relative performance decay of a model within the high-similarity band ($\Phi \geq 0.9$). It is defined as:

$$AD = \frac{P_{\text{train}} - P_{\Phi \geq 0.9}}{P_{\text{train}}}$$

where P_{train} is the average task success rate under training conditions, and $P_{\Phi \geq 0.9}$ is the average success rate in high-similarity test scenarios. The coefficient ranges from 0 to 1, with a lower value indicating better robustness to minor distribution shifts.

(2) Fluctuation Degree (FD): Measures the stability (inverse of dispersion) of performance within the high-similarity band, calculated from the coefficient of variation.

$$FD = \frac{\mu_P}{\mu_P + \sigma_P} \tag{23}$$

where μ_P and σ_P are the mean and standard deviation of the performance values $P(M, s)$ for all test scenarios s belonging to the high-similarity band $S_{\text{high-sim}}$.

(3) Generalization Capability Volume (GCV): A macro-level metric evaluating the model's overall adaptation across the entire test space, computed as the volume under the 3D performance landscape.

$$GCV = \sum_{i=1}^{I} \sum_{j=1}^{J} \overline{p}_{ij} \cdot \Delta\phi_i \cdot \Delta\psi_j \tag{24}$$

Here, I and J are the total numbers of bins along the similarity Φ and difficulty Ψ axes, respectively. $\overline{p}_{ij}$ is the average performance of scenarios within the two-dimensional(2D) bin indexed by (i, j), and $\Delta\phi_i$ and $\Delta\psi_j$ are the widths of the i-th similarity bin and j-th difficulty bin, respectively.

(4) Difficulty Sensitivity Index (DSI): Measures the average sensitivity of the model's performance to changes in scenario difficulty across different similarity levels.

$$DSI = \frac{1}{I} \sum_{i=1}^{I} |\beta_i| \tag{25}$$

where I is the number of similarity bins, and β_i is the slope of the linear regression line fitted to the performance values against difficulty levels Ψ for all scenarios within the i-th similarity bin.

(5) Generalization Stability Gradient (GSG): Quantifies the model's comprehensive sensitivity to joint changes in similarity and difficulty, represented by the average steepness (gradient magnitude) of the performance landscape.

$$GSG = \frac{1}{I \cdot J} \sum_{i=1}^{I} \sum_{j=1}^{J} ||\nabla P(i,j)|| \tag{26}$$

Here, I and J are the total numbers of bins along the Φ and Ψ axes. $\nabla P(i,j)$ denotes the gradient vector of the interpolated performance landscape $P = f(\Phi, \Psi)$ at the grid point corresponding to the (i,j) bin, and $||\cdot||$ denotes the vector norm (magnitude).

(6) Proportion of Effective Generalization Space (PEGS): Defines the model's reliable operational boundary by calculating the proportion of the test space where performance exceeds a predefined safety threshold $P_{\min}$.

$$PEGS = \frac{\sum_{i=1}^{I} \sum_{j=1}^{J} \mathbb{I}(\bar{P}_{ij} \geq P_{\min}) \cdot \Delta\phi_i \cdot \Delta\psi_j}{(\phi_{\max} - \phi_{\min})(\psi_{\max} - \psi_{\min})} \tag{27}$$

In this formula, $\mathbb{I}(\cdot)$ is the indicator function (returns 1 if the condition is true, 0 otherwise). $\bar{P}_{ij}$ is the average performance within bin (i,j). $\Delta\phi_i$ and $\Delta\psi_j$ are the bin widths. $\phi_{\max}, \phi_{\min}$ and $\psi_{\max}, \psi_{\min}$ are the maximum and minimum values of the similarity and difficulty dimensions in the test set, respectively.

(7) Surface Centroid Value (SCV): Evaluates the model's baseline performance under typical test conditions, defined as the value on the performance landscape at the average similarity and average difficulty level:

$$S_{CV} = Z(\Phi_{med}, \Psi_{med})$$

where Z is the performance response surface, Φ_{med} and Ψ_{med} are the median difficulty and similarity levels across the test set, respectively. A higher S_{CV} indicates reliable performance under moderately unfamiliar and typically difficult scenarios, reflecting the algorithm's core generalization strength.

This 3D dynamic evaluation methodology completes the framework, transforming raw scenario-level results into a structured, interpretable, and multi-dimensional assessment of a UGV decision-making system's generalization and robustness.

4. Experiment

To systematically validate the proposed dynamic evaluation framework and demonstrate its diagnostic capability, we conducted experiments based on a customized 2D continuous-space simulation environment. This environment faithfully simulates the kinematics and dynamics of the UGVs but does not currently integrate a perception module. Consequently, the scenario similarity quantification in this experimental validation focuses on those aspects of the framework that are independent of perceptual input, serving as a proof of concept for the core quantification mechanism. The primary objective of this study is to validate the framework's diagnostic power, rather than to directly compete with algorithmic paradigms. Therefore, we intentionally include both traditional and learning-based planners to demonstrate how the framework can characterize and differentiate their distinct behavioral patterns across the similarity-difficulty spectrum, even in a domain theoretically favorable to some methods.

4.1. Seed Scenario and Training

We designed a typical and comprehensive seed scenario for path planning, as shown in Figure 4, which encompasses core challenges such as obstacle avoidance, narrow passage traversal, and dynamic interaction. This serves as a classic testbed for evaluating the generalization capability of navigation algorithms [42,43]. This seed scenario acts as the sole training environment for all learning-based algorithms and as the distributional center for generating diverse test scenarios. Five navigation algorithms, comprising three learning-based methods (A2C, PPO, SAC) and two traditional planners (DWA, RRT), are evaluated in this study. Accordingly, all evaluated algorithms were deployed and finely tuned until stable performance was achieved within this seed scenario [44,45]. Upon reaching this stable state, their average success rates in the seed scenario were recorded as follows: A2C: 71%, DWA: 100%, PPO: 88%, RRT: 100%, and SAC: 24%. The specific structured parameters of the scenario are detailed in Table 2.

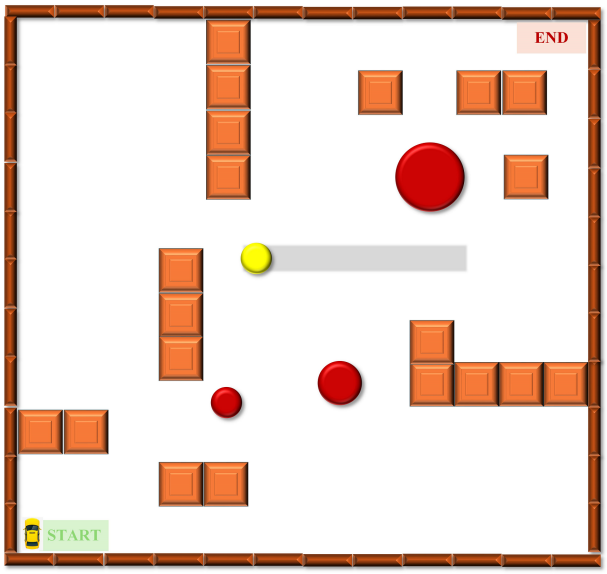

Figure 4. The benchmark seed scenario for algorithm training, featuring core navigation challenges like obstacle avoidance, narrow passage traversal, and dynamic interaction. The yellow circle represent dynamic obstacles, and the red circles represent static obstacles.

The two semantic rules chosen, narrow passage existence and dynamic trajectory conflict, represent core high-level challenges in navigation, allowing the semantic layer to contribute meaningfully to similarity assessment within the scope of this path-planning study.

4.2. Test Scenario Database Generation and Distribution

We constructed a test set that covers a wide range of similarity. It contains 829 candidate scenarios. To create them, we applied large random perturbations to the key parameters of a seed scenario [46]. These perturbations were sufficiently large to generate low-similarity scenarios ($\Phi < 0.3$), which have visibly distinct layouts and interactions. This approach simulates a gradual progression into out-of-distribution conditions for path planning.

The filtered test scenarios were then divided into two dedicated subsets for different analytical purposes. First, a high-similarity test set of 82 scenarios ($\Phi \geq 0.9$) was constructed for in-depth robustness analysis. The number of scenarios in each similarity sub-interval within this band is: 10 ($[1.0, 0.98]$), 12 ($[0.98, 0.96]$), 20 ($[0.96, 0.94]$), 20 ($[0.94, 0.92]$), and 20 ($[0.92, 0.90]$). Second, a larger generalization test set of 317 scenarios was formed to ensure broad and relatively uniform coverage across the quantized similarity–difficulty space for

comprehensive generalization analysis. The distribution of these 317 scenarios across the two-dimensional similarity–difficulty space is detailed in Table 3, confirming that our test database achieves good coverage across both dimensions and provides a solid foundation for subsequent fine-grained evaluation.

Table 3. Distribution of the 317-scenario generalization test set across the quantized similarity–difficulty space.

Similarity Interval	Difficulty Level										Total
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	
$[0.9, 1.0]$	6	5	4	5	3	4	3	2	1	0	33
$[0.8, 0.9]$	5	4	4	5	4	2	3	1	0	0	28
$[0.7, 0.8]$	5	4	5	3	5	3	3	2	1	0	31
$[0.6, 0.7]$	3	4	4	4	4	3	4	2	1	1	30
$[0.5, 0.6]$	2	3	5	4	5	4	3	4	1	1	32
$[0.4, 0.5]$	1	3	6	6	4	3	3	3	5	1	35
$[0.3, 0.4]$	2	1	2	3	5	3	3	4	4	4	31
$[0.2, 0.3]$	0	1	1	4	2	5	4	2	5	6	30
$[0.1, 0.2]$	0	0	2	2	3	4	5	6	6	7	35
$[0, 0.1]$	0	1	2	2	4	6	5	3	4	5	32
Total	24	26	35	38	39	37	36	29	28	25	317

4.3. Implementation of Similarity and Difficulty Computation

(1) Similarity Computation.
Following the methodology outlined in Section 3.1, we computed the similarity score $\Phi(s)$ for each test scenario relative to the training distribution. The resulting similarity values were then discretized by evenly dividing the range $[0.2, 1.0]$ into 10 intervals for subsequent analysis.

(2) Difficulty Annotation.
LLM Jury (A Priori Reasoning): In this study, we set $K_1 = 4$ and called the following specific model endpoints through their official APIs, employing a secure API relay service to ensure stable connectivity: gpt-5-mini-2025-08-07, gemini-3-pro-preview, DeepSeek-V3.2-Exp, and qwen3-max. All calls were made between 1 November and 15 December 2025.
The relay service functioned solely as a transparent intermediary and did not affect the models' input-output behavior. Each test scenario was automatically converted into a structured textual description (Figure 3) as a unified input for the LLMs. Each LLM then outputs a difficulty level between 1 and 10 along with a self-reported confidence score.

Proxy Model Jury (A Posteriori Performance): To obtain an independent, objective, and diverse performance baseline, we set $K_2 = 8$ and selected the following set of algorithms, which includes the five algorithms under evaluation as well as three additional benchmark methods: A* (graph search), Artificial Potential Field (APF), and Deep Q-Network (DQN). Each representing a different algorithmic paradigm, to constitute the proxy model jury [47,48]. Each algorithm was independently executed in each scenario, and its task success rate was used to compute an initial difficulty score. The final difficulty level for this jury is the mean of the scores from the eight algorithms, and its confidence is determined by the overall standard deviation of these scores (see formulas in Section 3.2).

The outputs from the two juries were fused via the weighted consensus mechanism described in Section 3.2.2 to generate the final integer difficulty level $\Psi(s) \in [1, 10]$ for each scenario. The high inter-jury consistency of the final annotations validates the reliability of the process.

5. Results

This chapter systematically evaluates the performance of the five algorithms (A2C, DWA, PPO, RRT, and SAC) within the proposed 3D evaluation framework of "similarity–difficulty–success rate," based on the experimental setup described in Section 4. We first analyze the robustness of each algorithm within the high-similarity band ($\Phi \geq 0.9$), followed by a feature analysis using 3D visualization. Finally, a multi-dimensional quantitative comparison is conducted, leading to practical recommendations for applicability.

5.1. Robustness Analysis

Following the evaluation framework established in Section 3, this section focuses on analyzing the robustness of each algorithm in high-similarity scenarios. We selected test scenarios with similarity $\Phi \geq 0.9$ to form the high-similarity test set $S_{\text{high-sim}}$ (totaling 82 scenarios) and performed a comprehensive assessment across three levels: the distribution of task success rates, performance trends across similarity sub-intervals, and core quantitative metrics.

To provide an intuitive comparison of algorithm performance within $S_{\text{high-sim}}$, Figure 5 presents a scatter plot of the task success rate for each algorithm across all test scenarios.

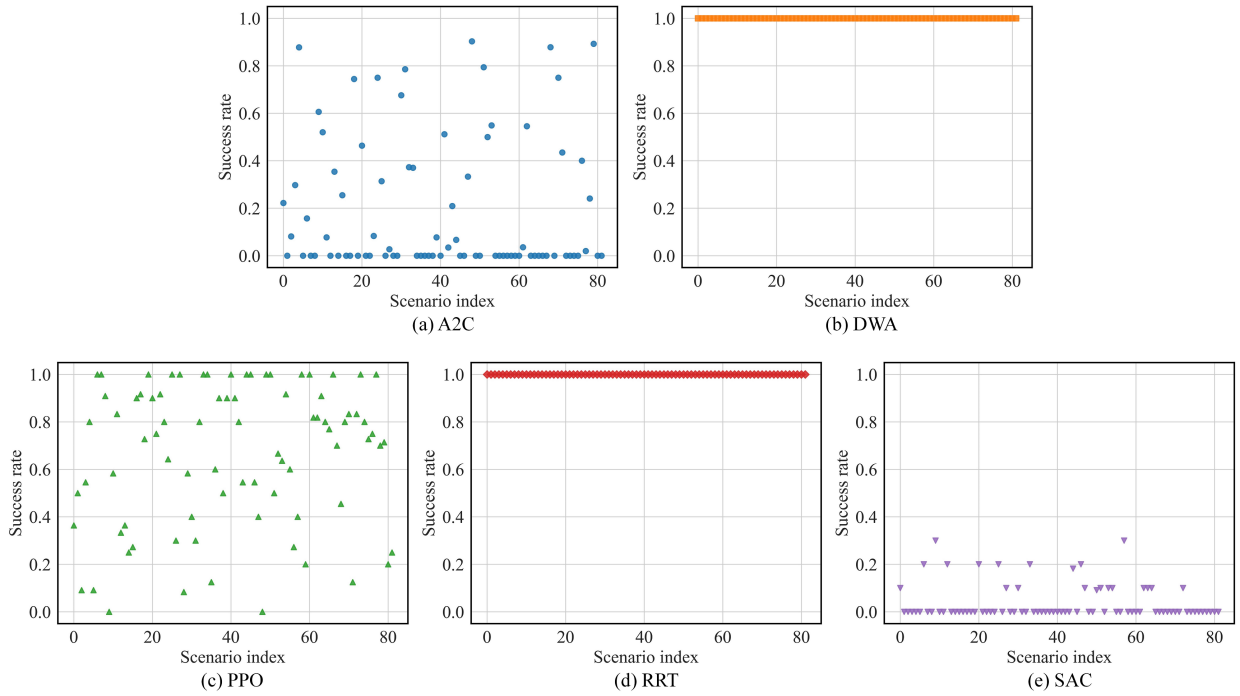

Figure 5. Performance distribution within the high-similarity band ($\Phi \geq 0.9$) reveals stark differences in robustness between traditional and learning-based algorithms.

DWA and RRT maintain compact, high-success-rate clusters (Figure 5), demonstrating near-perfect robustness. This stems from their model-based, non-learning nature, which makes performance inherently stable and independent of training data variations.

In contrast, the point clusters for A2C and SAC are predominantly located in lower success-rate regions, with numerous instances of zero success rate, reflecting unstable and poor performance within this band.

To further investigate the trend of performance with subtle variations in similarity, we subdivided the high-similarity band $[0.9, 1.0]$ into five sub-intervals and calculated the average success rate for each algorithm within them, as shown in Figure 6.

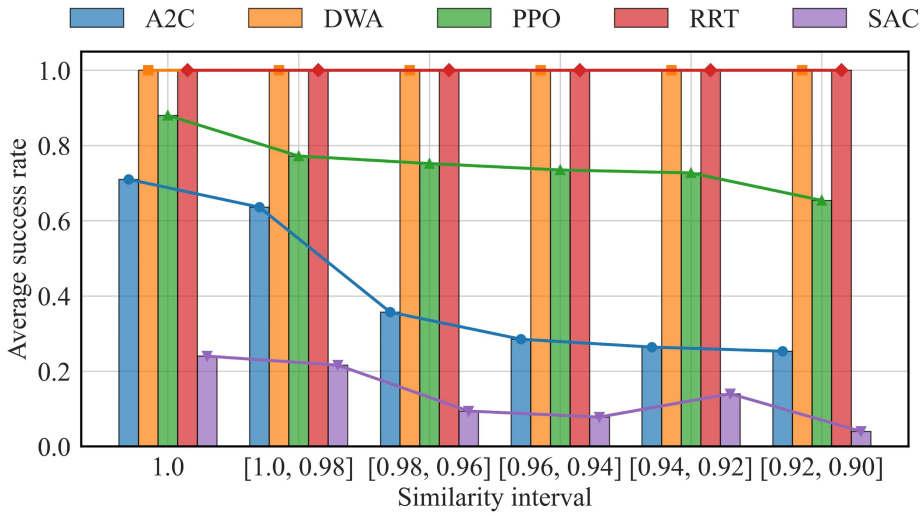

Figure 6. Similarity-dependent performance decay within the high-similarity band quantifies algorithmic robustness to minor perturbations.

As observed in Figure 6, as the similarity decreases gradually from 1.0 to 0.9, the average success rate for all algorithms shows a declining trend, which is expected. Among them, DWA and RRT maintain the highest and most stable success rates across all subintervals. The decline for A2C and SAC is particularly pronounced, with performance degradation accelerating notably when similarity falls below 0.98. The decline for PPO is relatively moderate but also shows a marked drop when similarity falls below 0.92.

For a precise quantitative assessment of algorithm robustness, we calculated two core metrics: Attenuation Degree (AD) and Fluctuation Degree (FD). The results are presented in Table 4.

Table 4. Quantitative assessment of in-distribution robustness using attenuation degree (AD) and fluctuation degree (FD) metrics.

Algorithm	Avg. Success Rate in Seed Scenarios	Avg. Success Rate in High-Similarity Band	Attenuation Degree ↓	Fluctuation Degree ↓
A2C	71.00%	35.934%	49.389%	1.437
DWA	100.00%	100.00%	0	0
PPO	88.00%	72.865%	17.199%	0.450
RRT	100.00%	100.00%	0	0
SAC	24.00%	11.368%	52.633%	1.912

↓ indicates cost indicators.

(1) DWA and RRT demonstrated near-perfect robustness, with both AD and FD being zero, indicating their performance is almost unaffected by minor perturbations within high-similarity scenarios.
(2) PPO exhibited a more noticeable attenuation degree, suggesting its performance is relatively susceptible to subtle scenario changes. However, its relatively lower fluctuation index reflects a certain level of stability in its performance.
(3) Both AD and FD were significantly higher for A2C and SAC, indicating they are highly sensitive to minor perturbations and exhibit poor robustness.

A comprehensive analysis indicates the following robustness ranking for the five algorithms within the high-similarity band: DWA = RRT $\succ$ PPO $\succ$ A2C $\succ$ SAC. These quantified values for AD and FD directly address Objective (4) from Section 2.4, providing decoupled metrics that specifically assess the ID Robustness of each algorithm.

5.2. Generalization Analysis

5.2.1. Qualitative Patterns from 3D Performance Landscapes

Based on the method proposed in Section 3.3, we constructed the performance landscape of each algorithm in the 3D "success rate–similarity–difficulty" space (Figure 7), and intuitively revealed their performance trends by analyzing graphical characteristics (Table 5).

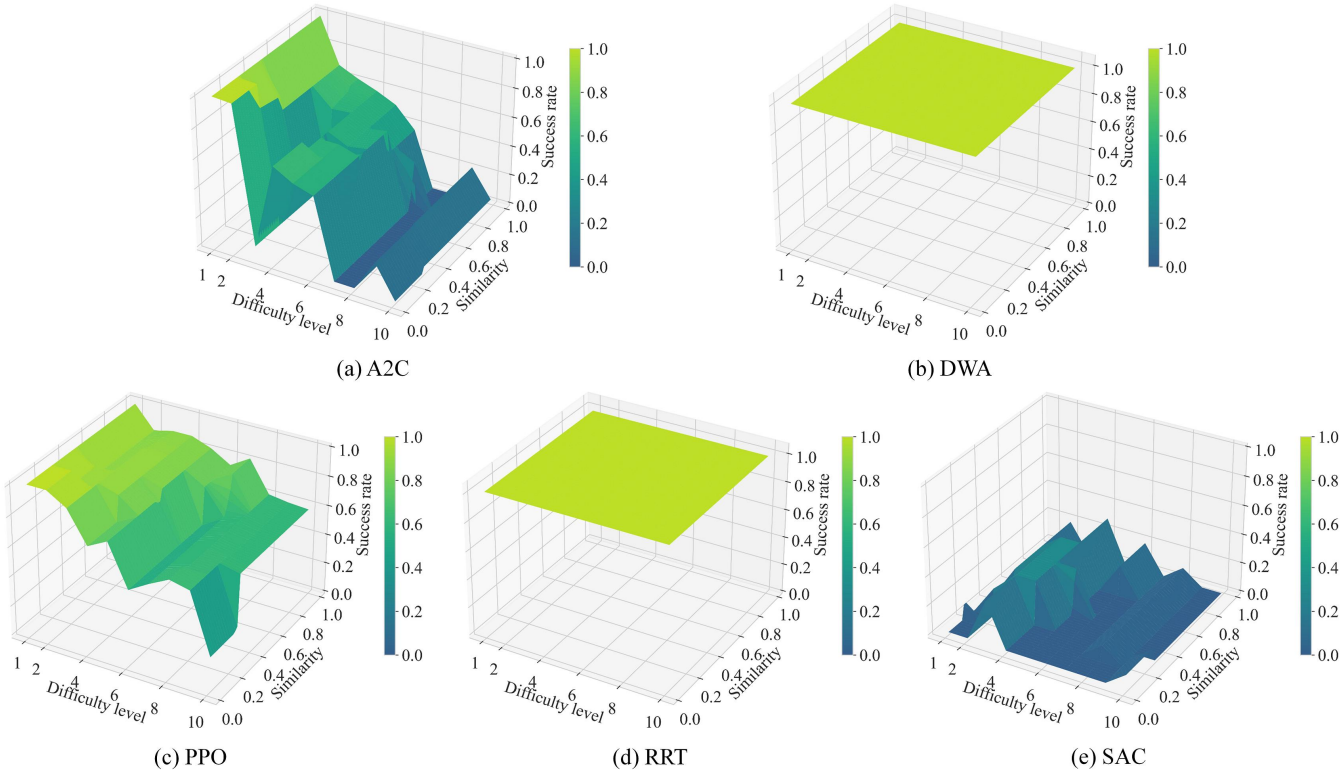

Figure 7. Diagnostic 3D performance landscapes illustrate distinct generalization patterns across the similarity-difficulty space.

Table 5. Qualitative analysis of the 3D performance landscapes reveals characteristic generalization.

Graphical Characteristic Dimension	A2C	PPO	DWA/RRT	SAC
Surface Flatness	Extremely steep, highly fluctuant	Sloped, relatively smooth	Very flat	Highly undulating
Sensitivity to Similarity	Very high	High	Very low	Present but non-monotonic
Sensitivity to Difficulty	Very high	High	Very low	Present but non-monotonic
Generalization Pattern	Fragile yet unpredictable	Predictable monotonic decay	Robust	No discernible pattern

Collectively, the performance landscapes and their graphical characteristics (Figure 7, Table 5) highlight several distinct generalization patterns, which can be summarized as follows:

(1) For learning-based algorithms like A2C, PPO, and SAC, performance strongly correlates with similarity. Traditional planners such as DWA and RRT show no such sensitivity. This difference highlights the core challenge of distribution shift in data-driven methods.

(2) PPO has a surface shown in Figure 7. It is smooth and declining. Additionally, this surface illustrates a typical generalization decay. This decay pertains to a policy that is optimized for in-distribution reward. The severe fluctuations of the A2C surface expose potential issues of high variance and instability in certain on-policy or optimization methods.

(3) SAC's non-monotonic, 'arched' performance landscape (Figure 7) and low sensitivity metrics may arise from its maximum entropy objective, which encourages exploration and behavioral diversity at the potential cost of peak in-distribution robustness but can lead to more gradual degradation under distribution shift.

These distinct patterns revealed by the 3D landscape achieve Objective (3), enabling a diagnostic visualization of model behavior across the similarity–difficulty space that static metrics cannot provide.

5.2.2. Quantitative Comparison Using Interpretable Metrics

To precisely evaluate the generalization performance of each algorithm under variations in both similarity and difficulty, we further employed the five core metrics defined in Section 3.3.3 for quantitative analysis. The results are presented in Table 6, and a multi-dimensional performance comparison is illustrated in Figure 8.

Table 6. Comprehensive quantitative evaluation of generalization using five metrics extracted from the performance landscape.

Algorithm	GCV ↑	DSI ↓	GSG ↓	PEGS ↑	SCV ↑
A2C	0.473	0.398	1.547	0.143	0.375
DWA	1	0	0	1	1
PPO	0.746	0.166	0.907	0.227	0.734
RRT	1	0	0	1	1
SAC	0.036	0.035	0.381	0	0.162

↓ indicates cost indicators; ↑ indicates benefit indicators.

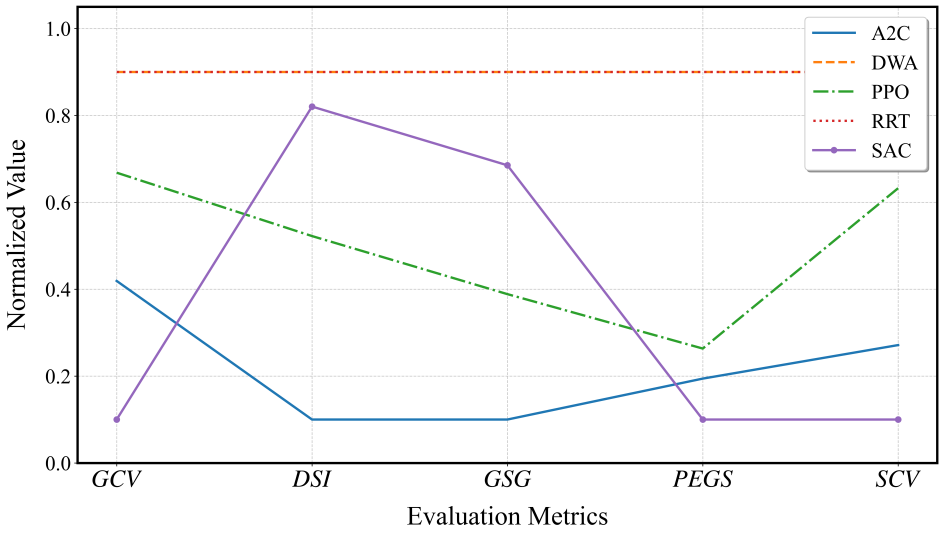

Figure 8. Multidimensional Algorithm Comparison Using Five Diagnostic Metrics.

The integrated analysis of the fine-grained metrics (Table 6) and the multidimensional visualization (Figure 8, which is based on normalized data) elucidates the distinct algorithmic profiles revealed by our framework.

(1) DWA and RRT have DSI and GSG values equal to zero, and GCV, PEGS, and SCV are 1, indicating near-ideal generalization stability and full coverage capability, with extremely slow performance decay under distribution shift.
(2) PPO performs moderately on most metrics, with its GCV, PEGS, and SCV significantly better than those of A2C and SAC.
(3) A clear contrast exists between A2C and SAC. In GCV, PEGS, and SCV, A2C substantially outperforms SAC, whose notably low GCV and zero PEGS indicate widespread failure and a weak safety boundary. In DSI and GSG, however, SAC exhibits lower sensitivity and a more gradual performance decay, a pattern reflected in its more balanced radar profile. This contrast underscores that the "height" of performance and the "slope" of its decay represent two distinct and decoupled dimensions of generalization behavior.

The multi-faceted profile from these five metrics (GCV, DSI, GSG, PEGS, SCV) collectively fulfills Objective (4) by offering a comprehensive, quantitative characterization of each algorithm's OOD Generalization capability and its operational boundaries.

5.3. Summary and Practical Implications

Therefore, if the primary objective is to achieve high-performance coverage across scenarios, the ranking based on core performance metrics is: DWA = RRT $\succ$ PPO $\succ$ A2C $\succ$ SAC. This order offers a practical guideline for algorithm selection in safety-critical applications. These results, along with the revealed decoupling between performance level and decay stability, demonstrate the diagnostic power of our framework in exposing clear divergences in both robustness and generalization capabilities (divergences that are masked by conventional aggregate metrics). To further elucidate this diagnostic value, the following discussion section will analyze these findings in depth and systematically contrast them with conventional evaluation paradigms.

6. Discussion

6.1. Diagnostic Advantage over Conventional Evaluation Paradigms

To underscore the diagnostic value of our framework, we contrast it with the conventional evaluation paradigms that rely on monolithic aggregate metrics. When evaluated solely on average success rate across the entire test set of 317 scenarios, the algorithms achieve: A2C: 60%, PPO: 72%, SAC: 58%, with DWA and RRT both at 100%. This coarse-grained ranking masks critical behavioral distinctions. For instance, while A2C and SAC show similar aggregate scores, our fine-grained metrics (Table 6) reveal starkly different failure modes: SAC exhibits extreme sensitivity to minor perturbations (high AD and FD) and near-complete failure under distribution shift (GCV = 0.036, PEGS = 0), whereas A2C, though fragile, retains some capability in moderately novel scenarios (SCV = 0.375). Similarly, PPO's moderate aggregate score conceals its predictable, similarity-dependent decay—a pattern clearly captured by our 3D performance landscapes (Figure 7c). Thus, while traditional metrics offer a superficial performance summary, our framework disentangles robustness from generalization, exposing root causes of degradation and providing actionable insights for safety-aware algorithm selection and optimization.

6.2. Interplay Between Robustness and Generalization

By evaluating robustness within the high-similarity band ($\Phi \geq 0.9$) and generalization across the full similarity range, our framework reveals a nuanced relationship between

the two capabilities.The experimental results of our framework reveal a profound insight that transcends conventional evaluation: the robustness and generalization of an algorithm are not in a simple positive correlation. Instead, they may exhibit distinct, and at times even conflicting, characteristics. This conclusion is clearly evidenced by the comparative analysis between SAC and A2C.

As shown in Table 7, on the metrics specifically designed to measure in-distribution robustness (within the high-similarity band), SAC underperforms A2C. SAC exhibits higher performance attenuation (AD: 52.633%) and fluctuation (FD: 1.912), indicating greater sensitivity to minor perturbations within the training distribution and poorer stability. However, SAC demonstrates a markedly different pattern on metrics that characterize full-space generalization behavior. Its Difficulty Sensitivity Index (DSI: 0.035) and Generalization Stability Gradient (GSG: 0.381) are significantly lower than those of A2C (DSI: 0.398; GSG: 1.547). This implies that SAC's performance degrades more gradually with increasing scenario difficulty and distributional shift, lacking the sharp, predictable decline observed in A2C. This "low-sensitivity" trait is visually corroborated by its 3D performance landscapes in Figure 7(e), which, despite an overall lower position, displays a unique "arched" structure in the medium-similarity range. This suggests a potential, albeit underutilized, adaptive capacity in certain novel scenarios.

Table 7. Comparative analysis of key robustness and generalization metrics between A2C and SAC.

Evaluation Dimension	Metric	A2C	SAC	Revealed Relationship
Robustness High-Similarity	AD ↓	49.389%	52.633%	SAC exhibits more severe performance degradation, indicating weaker robustness.
	FD ↓	1.437	1.912	SAC shows greater performance volatility, indicating poorer stability.
Generalization Capability Full Space	DSI ↓	0.398	0.035	SAC is highly insensitive to changes in scenario difficulty.
	GSG ↓	1.547	0.381	SAC's response to distributional shift is more moderate overall.

↓ indicates cost indicators.

This finding aligns with a consensus in reinforcement learning. The consensus suggests a trade-off may exist between training performance and generalization capability. The underlying mechanism stems from core algorithmic designs. SAC is built upon the maximum entropy principle. It prioritizes exploration and behavioral diversity. This priority may sacrifice peak robustness near the training distribution. However, it encourages the policy to cover a broader behavioral space. Consequently, when faced with genuine distributional shifts, SAC exhibits a generalization pattern with lower sensitivity and potential adaptability. Conversely, algorithms like A2C, which prioritize stable convergence within the training environment, are more prone to overfitting. This leads to a steep generalization curve and high sensitivity to variations.

Therefore, the value of our framework lies not only in diagnostics but also in guidance. It demonstrates that for safety-critical real-world deployment, the choice of algorithm must be based on a rational trade-off informed by the core requirements of the task—whether

it demands extreme stability in known environments or broad adaptability to unknown variations—rather than blindly optimizing for a single aggregate metric.

6.3. Sensitivity Analysis of the High-Similarity Threshold

The threshold γ defines $S_{\text{high-sim}}$ for dedicated robustness evaluation. To examine the sensitivity of our findings to this parameter, we conducted additional analyses using $\gamma = 0.85$ and $\gamma = 0.95$. With $\gamma = 0.95$, $S_{\text{high-sim}}$ becomes more conservative, containing fewer scenarios. With $\gamma = 0.85$, the band expands to include more scenarios with moderate similarity. Crucially, the relative robustness ranking of the evaluated algorithms within $S_{\text{high-sim}}$ remained consistent: DWA and RRT consistently showed near-perfect stability, followed by PPO, then A2C and SAC. The key finding of a decoupled relationship between in-distribution robustness and out-of-distribution generalization behavior also held across these different γ values. While the absolute values of the metrics shifted slightly with the change in the evaluated scenario set, the comparative patterns and diagnostic conclusions were unchanged. This confirms that the core insights provided by our framework are robust to minor, reasonable variations in the choice of γ.

6.4. Limitations and Future Work

While the proposed framework demonstrates promising diagnostic capabilities, several limitations of this study should be noted.

First, as noted in the introduction, our experimental validation is conducted in a simplified 2D simulation environment focusing on the planning layer, and does not currently integrate a perception module. This choice was made to provide a clear proof-of-concept for the evaluation framework by decoupling it from the challenges of perceptual uncertainty. Consequently, the current results primarily speak to generalization in the decision-making space given perfect state information. A critical and necessary next step is to integrate raw sensory inputs (e.g., camera, LiDAR) to validate the framework's ability to diagnose generalization failures that stem from the perception module, thereby testing the full UGV autonomy stack under more realistic conditions [49].

Second, the difficulty annotation, though objective, is contingent on the chosen set of proxy models and LLMs. Exploring more efficient continuous difficulty metrics is an important direction.

Furthermore, future work will apply the framework to more challenging domains where traditional planners may struggle (e.g., environments with dense dynamic obstacles or complex local minima). This will further validate and extend the framework's diagnostic scope in scenarios that better highlight the adaptation challenges for learning-based systems.

Lastly, our current OOD validation tests performance across a continuous similarity spectrum generated via parameter perturbations. Including scenarios with fundamentally different topologies or semantics would test a stronger form of generalization. Our framework's similarity model is designed for such extension, and this remains an important future direction.

6.5. Practical Implications and Recommendations

Based on the above robustness and generalization analysis, we provide the following recommendations for the application scenarios of each algorithm:

(1) DWA and RRT, which demonstrated the best performance in this evaluation, exhibit strong robustness, high generalization stability, and full coverage capability. They are recommended for scenarios with extremely high safety requirements.

(2) PPO provides reliable performance in scenarios with moderate similarity and difficulty, but shows vulnerability under conditions of extreme low similarity or high difficulty. It is suitable for deployment environments where conditions are relatively controlled.
(3) Although A2C and SAC are learning-based algorithms, they exhibited poor robustness and generalization consistency in this task setup. We recommend further optimization of these algorithms before practical application.

6.6. Pathways to Real-World Experimentation

This diagnostic framework is ultimately intended for real-world deployment. Applying it to physical UGVs requires two main adjustments:

First, the similarity quantification module must process raw sensor data. This step needs to handle real-world perceptual noise and incomplete observations. Its goal is to reliably extract the four-layer scenario features under practical conditions.

Second, the difficulty annotation pipeline must adapt for onboard or digital-twin use. The key here is to maintain a consistent and practical objective difficulty assessment outside controlled simulation.

Real-world validation would take place in instrumented test fields. It would finally test the framework's ability to diagnose failures caused by real-world factors. These include perceptual noise, actuation dynamics, and complex environmental interactions not captured in simulation.

This step is crucial. It advances the framework toward real-world safety certification and defines true operational performance boundaries.

7. Conclusions

In conclusion, we have developed and validated a fine-grained dynamic evaluation framework that fundamentally shifts how generalization in UGVs is assessed. By moving from monolithic metrics to a diagnostic similarity–difficulty landscape, the framework introduces two key methodological pillars: a hierarchical, four-layer similarity quantification and a hybrid consensus mechanism for objective difficulty annotation. Together, they enable the construction of an interpretable 3D performance landscape.

The practical value of this framework is multifaceted. For algorithm development and optimization, the detailed performance landscape and the suite of seven metrics provide precise diagnostics. They can identify whether a failure mode stems from poor in-distribution robustness or a fundamental out-of-distribution generalization limit, guiding targeted improvements. For system deployment and safety certification, the framework offers actionable insights. Metrics like PEGS can directly inform the definition of a system's verified operational boundary. The clear decoupling of robustness and generalization supports informed, safety-aware algorithm selection tailored to specific deployment profiles, such as prioritizing extreme stability for controlled environments or broader adaptability for unknown terrain. Consequently, this work provides a step toward more rigorous, standardized, and interpretable benchmarking that is essential for certifiable autonomy in robotic systems. Building on these insights, a promising direction for future work involves integrating path quality metrics (e.g., path length, smoothness, and energy efficiency) into this diagnostic landscape to further discriminate algorithm performance beyond task success and extending the evaluation framework to 3D navigation spaces.

Author Contributions: Conceptualization, Z.D. and Y.G. and J.Y. and X.T. ;methodology, Z.D. and Y.G. and J.Y. and X.T.; software, Z.D. and Y.G. and J.Y. and X.T. and W.X. and M.L.; validation, Z.D. and Y.G. and J.Y. and X.T. and W.X. and M.L.; formal analysis, Z.D. and Y.G. and J.Y. and X.T. and W.X. and M.L.; investigation, Z.D. and Y.G. and J.Y. and X.T. and W.X. and M.L.; data curation, Z.D. and Y.G. and J.Y. and X.T. and W.X. and M.L.; writing—original draft preparation, Z.D. and Y.G. and J.Y. and X.T.; writing—review and editing, Z.D. and Y.G. and J.Y. and X.T.; supervision, Z.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

UGV	Unmanned Ground Vehicle
3D	three-dimensional
2D	two-dimensional
OOD	Out-of-Distribution
ID	In-Distribution
LLM	Large Language Model
AHP	Analytic Hierarchy Process
AD	Attenuation Degree
FD	Fluctuation Degree
GCV	Generalization Capability Volume
DSI	Difficulty Sensitivity Index
GSG	Generalization Stability Gradient
PEGS	Proportion of Effective Generalization Space
SCV	Surface Centroid Value

References

1. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444.
2. Deng, Y.; Chow, A.H.; Yan, Y.; Su, Z.; Zhou, Z.; Kuo, Y.H. Hierarchical production control and distribution planning under retail uncertainty with reinforcement learning. *Int. J. Prod. Res.* **2025**, *63*, 4504–4522.
3. Dong, X.; Hua, Y.; Zhou, Y.; Ren, Z.; Zhong, Y. Theory and experiment on formation-containment control of multiple multirotor unmanned aerial vehicle systems. *IEEE Trans. Autom. Sci. Eng.* **2018**, *16*, 229–240.
4. Yu, Y.; Li, J.; Wu, T. A Group Target Tracking Method for Unmanned Ground Vehicles Based on Multi-Ellipse Shape Modeling. *Drones* **2025**, *9*, 620.
5. Dong, Z.; Yang, J.; Yuan, R.; Su, G.; Lei, M. A Game-Theoretic Kendall's Coefficient Weighting Framework for Evaluating Autonomous Path Planning Intelligence. *Automation* **2025**, *6*, 85.
6. Brunke, L.; Greeff, M.; Hall, A.W.; Yuan, Z.; Zhou, S.; Panerati, J.; Schoellig, A.P. Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annu. Rev. Control. Robot. Auton. Syst.* **2022**, *5*, 411–444.
7. Hendrycks, D.; Basart, S.; Mu, N.; Kadavath, S.; Wang, F.; Dorundo, E.; Desai, R.; Zhu, T.; Parajuli, S.; Guo, M.; et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 8340–8349.
8. Li, Y.; Yuan, W.; Zhang, S.; Yan, W.; Shen, Q.; Wang, C.; Yang, M. Choose your simulator wisely: A review on open-source simulators for autonomous driving. *IEEE Trans. Intell. Veh.* **2024**, *9*, 4861–4876.
9. Gulino, C.; Fu, J.; Luo, W.; Tucker, G.; Bronstein, E.; Lu, Y.; Harb, J.; Pan, X.; Wang, Y.; Chen, X.; et al. Waymax: An accelerated, data-driven simulator for large-scale autonomous driving research. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 7730–7742.
10. Zhou, K.; Liu, Z.; Qiao, Y.; Xiang, T.; Loy, C.C. Domain generalization: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 4396–4415.
11. Wang, J.; Lan, C.; Liu, C.; Ouyang, Y.; Qin, T.; Lu, W.; Chen, Y.; Zeng, W.; Yu, P.S. Generalizing to unseen domains: A survey on domain generalization. *IEEE Trans. Knowl. Data Eng.* **2022**, *35*, 8052–8072.

12. Gong, L.; Zhang, Y.; Xia, Y.; Zhang, Y.; Ji, J. SDAC: A multimodal synthetic dataset for anomaly and corner case detection in autonomous driving. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 26–27 February 2024; Volume 38, pp. 1914–1922.
13. Dong, Y.; Kang, C.; Zhang, J.; Zhu, Z.; Wang, Y.; Yang, X.; Su, H.; Wei, X.; Zhu, J. Benchmarking robustness of 3d object detection to common corruptions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 1022–1032.
14. Li, J.; Zong, R.; Wang, Y.; Deng, W. Complexity Evaluation for Urban Intersection Scenarios in Autonomous Driving Tests: Method and Validation. *Appl. Sci.* **2024**, *14*, 10451.
15. Yang, S.; Wang, C.; Zhang, Y.; Yin, Y.; Huang, Y.; Li, S.E.; Chen, H. Quantitative representation of autonomous driving scenario difficulty based on adversarial policy search. *Research* **2025**, *8*, 0575.
16. Ayerdi, J.; Iriarte, A.; Valle, P.; Roman, I.; Illarramendi, M.; Arrieta, A. Marmot: Metamorphic runtime monitoring of autonomous driving systems. *ACM Trans. Softw. Eng. Methodol.* **2024**, *34*, 1–35.
17. Dong, Z. Dynamic assessment of threats to cluster targets inlowaltitude multi-domain battlefields. *J. Tsinghua Univ. (Sci. Technol.)* **2024**, *64*, 1380–1390.
18. Xiong, Z.; Chen, S. A multi-fidelity approach for reliability-based risk assessment of single-vehicle crashes. *Accid. Anal. Prev.* **2024**, *195*, 107391.
19. Islam, M.; Li, Z.; Glocker, B. Robustness stress testing in medical image classification. In Proceedings of the International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, Vancover, BC, Canada, 12 October 2023; Springer: Berlin/Heidelberg, Germany, 2023; pp. 167–176.
20. Dong, Z.; Yang, J.; Su, G.; Guo, Y.; Lei, M.; Liu, X.; Shi, Y. Dynamic Evaluation of Learning Internalization Capability in Unmanned Ground Vehicles via Time Series Analysis. *Drones* **2026**, *10*, 44.
21. Koh, P.W.; Sagawa, S.; Marklund, H.; Xie, S.M.; Zhang, M.; Balsubramani, A.; Hu, W.; Yasunaga, M.; Phillips, R.L.; Gao, I.; et al. Wilds: A benchmark of in-the-wild distribution shifts. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; PMLR: London, UK, pp. 5637–5664.
22. Singh, A.; Sarangmath, K.; Chattopadhyay, P.; Hoffman, J. Benchmarking low-shot robustness to natural distribution shifts. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 16232–16242.
23. Dziugaite, G.K.; Drouin, A.; Neal, B.; Rajkumar, N.; Caballero, E.; Wang, L.; Mitliagkas, I.; Roy, D.M. In search of robust measures of generalization. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 11723–11733.
24. Hu, Y.; Fu, J.; Wen, G.; Lv, Y.; Ren, W. Distributed entropy-regularized multi-agent reinforcement learning with policy consensus. *Automatica* **2024**, *164*, 111652.
25. WANG, H.; YE, X.; DONG, Z. Residual network-based stacked vector quantized autoencoder. *J. Tsinghua Univ. (Sci. Technol.)* **2025**, *65*, 2259–2268.
26. Wang, K.; Shen, C.; Li, X.; Lu, J. Uncertainty quantification for safe and reliable autonomous vehicles: A review of methods and applications. *IEEE Trans. Intell. Transp. Syst.* **2025**, *26*, 2880–2896.
27. Guerrero-Sevilla, D.; Gonzalez-de Soto, M.; Del Pozo, S.; Martín-Jiménez, J.A.; Rodríguez-Gonzálvez, P.; González-Aguilera, D. Enhancing Overtaking Safety with Mobile LiDAR Systems: Dynamic Analysis of Road Visibility. *Remote Sens.* **2025**, *17*, 2948.
28. Poursaeed, O.; Jiang, T.; Yang, H.; Belongie, S.; Lim, S.N. Robustness and generalization via generative adversarial training. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 15711–15720.
29. Sünderhauf, N.; Brock, O.; Scheirer, W.; Hadsell, R.; Fox, D.; Leitner, J.; Upcroft, B.; Abbeel, P.; Burgard, W.; Milford, M.; et al. The limits and potentials of deep learning for robotics. *Int. J. Robot. Res.* **2018**, *37*, 405–420.
30. Freiesleben, T.; Grote, T. Beyond generalization: A theory of robustness in machine learning. *Synthese* **2023**, *202*, 109.
31. Yang, J.; Zhou, K.; Li, Y.; Liu, Z. Generalized out-of-distribution detection: A survey. *Int. J. Comput. Vis.* **2024**, *132*, 5635–5662.
32. Vysotska, O.; Bogoslavskyi, I.; Hutter, M.; Stachniss, C. Adaptive thresholding for sequence-based place recognition. In Proceedings of the 42nd IEEE International Conference on Robotics and Automation (ICRA 2025), Atlanta, GA, USA, 19–23 May 2025.
33. Zhang, X.; Zhao, W.; Sun, Y.; Sun, J.; Shen, Y.; Dong, X.; Yang, Z. Testing automated driving systems by breaking many laws efficiently. In Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis, Seattle, WA, USA, 17–21 July 2023; pp. 942–953.
34. Wang, D.; Devin, C.; Cai, Q.Z.; Krähenbühl, P.; Darrell, T. Monocular plan view networks for autonomous driving. In Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 3–8 November 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 2876–2883.
35. Kendall, A.; Gal, Y.; Cipolla, R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7482–7491.

36. Gao, M.; Hu, X.; Yin, X.; Ruan, J.; Pu, X.; Wan, X. Llm-based nlg evaluation: Current status and challenges. *Comput. Linguist.* **2025**, *51*, 661–687.
37. Szymanski, A.; Ziems, N.; Eicher-Miller, H.A.; Li, T.J.J.; Jiang, M.; Metoyer, R.A. Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks. In Proceedings of the 30th International Conference on Intelligent User Interfaces, Cagliari Italy, 24–27 March 2025; pp. 952–966.
38. Li, Y.; Zhang, S.; Wu, R.; Huang, X.; Chen, Y.; Xu, W.; Qi, G.; Min, D. MATEval: A multi-agent discussion framework for advancing open-ended text evaluation. In Proceedings of the International Conference on Database Systems for Advanced Applications, Gifu, Japan, 2–5 July 2024; Springer: Berlin/Heidelberg, Germany, 2024; pp. 415–426.
39. Yu, N.P.; Liu, C.C.; Price, J. Evaluation of market rules using a multi-agent system method. *IEEE Trans. Power Syst.* **2009**, *25*, 470–479.
40. Li, Y.; Zhang, Y.; Li, X.; Sun, C. Regional multi-agent cooperative reinforcement learning for city-level traffic grid signal control. *IEEE/CAA J. Autom. Sin.* **2024**, *11*, 1987–1998.
41. Rudin, C.; Chen, C.; Chen, Z.; Huang, H.; Semenova, L.; Zhong, C. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Stat. Surv.* **2022**, *16*, 1–85.
42. Haarnoja, T.; Zhou, A.; Abbeel, P.; Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; PMLR: London, UK, 2018; pp. 1861–1870.
43. Kiran, B.R.; Sobh, I.; Talpaert, V.; Mannion, P.; Al Sallab, A.A.; Yogamani, S.; Pérez, P. Deep reinforcement learning for autonomous driving: A survey. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 4909–4926.
44. Mendes, P.; Batista, P.; Oliveira, P.; Silvestre, C. Cooperative decentralized navigation algorithms based on bearing measurements for arbitrary measurement topologies. *Ocean Eng.* **2023**, *270*, 113564.
45. Gao, G.; Lu, J.; Guan, W. CE-Bi-RRT*: Enhanced Bidirectional RRT* with Cooperative Expansion Strategy for Autonomous Drone Navigation. *Drones* **2025**, *9*, 831.
46. Pinto, L.; Davidson, J.; Gupta, A. Supervision via competition: Robot adversaries for learning tasks. In Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, 29 May–3 June 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1601–1608.
47. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533.
48. Jaderberg, M.; Czarnecki, W.M.; Dunning, I.; Marris, L.; Lever, G.; Castaneda, A.G.; Beattie, C.; Rabinowitz, N.C.; Morcos, A.S.; Ruderman, A.; et al. Human-level performance in 3D multiplayer games with population-based reinforcement learning. *Science* **2019**, *364*, 859–865.
49. Philion, J.; Fidler, S. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 194–210.