automation

MDPI

Article

A Game-Theoretic Kendall's Coefficient Weighting Framework for Evaluating Autonomous Path Planning Intelligence

Zewei Dong [1,*], Jingxuan Yang [1], Runze Yuan [1], Guangzhen Su [1,2] and Ming Lei [1]

[1] Department of Automation, Tsinghua University, Beijing 100084, China; yangjx20@mails.tsinghua.edu.cn (J.Y.); yuanrz@mail.tsinghua.edu.cn (R.Y.); 1155248203@link.cuhk.edu.hk (G.S.); leim24@mails.tsinghua.edu.cn (M.L.)
[2] Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong, Hong Kong SAR, China
* Correspondence: dzw22@mails.tsinghua.edu.cn

Abstract

Accurately evaluating the intelligence of autonomous path planning remains challenging, primarily due to the interdependencies among evaluation metrics and the insufficient integration of subjective and objective weighting methods. This paper proposes Game-Theoretic Kendall's Coefficient (GTKC) weighting framework for evaluating autonomous path planning intelligence. The framework specifies a safety–efficiency–comfort metric system with observable, reproducible, and quantifiable metrics. To account for intermetric dependence, subjective weights are elicited via an improved Analytic Network Process (ANP), while objective weights are derived using the CRITIC method to capture contrast intensity and intercriteria conflict. The credibility of the subjective and objective weights is evaluated using Kendall's coefficient and the coefficient of variation, respectively. Subsequently, based on the principle that higher credibility should receive greater weight, a game-theoretic optimization model is employed to dynamically derive optimal combination coefficients. Experimental results on three case scenarios demonstrate that the GTKC framework significantly outperforms existing weighting approaches in terms of effectiveness (achieving a lowest Mean Absolute Error (MAE) of 0.15 and a perfect Spearman's correlation coefficient ($\bar{\rho} = 1.0$) with ground-truth rankings) , stability (Mean Standard Deviation (MSD) = 0.023), and ranking consistency (Kendall's coefficient W = 0.924). These findings validate GTKC as a theoretically grounded and practically robust mechanism that explicitly models metric interdependencies and integrates expert knowledge with empirical evidence, enabling reliable and reproducible evaluation of autonomous path planning intelligence.

Keywords: autonomous path planning; intelligence evaluation; combination weighting; Kendall's coefficient; game-theoretic

Academic Editor: Shuai (Steven) Li

Received: 17 October 2025
Revised: 18 November 2025
Accepted: 20 November 2025
Published: 2 December 2025

Citation: 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. https://doi.org/10.3390/automation6040085

1. Introduction

Autonomous path planning systems (APPSs) play a critical role in the unmanned ground vehicle (UGV) decision-making procedure [1–5]. APPS is primarily manifested in their capability to recognize traffic signals and obstacles [6], understand traffic rules and dynamic environmental changes [4,7], and dynamically adjust their operational states to select safe, comfortable, energy-efficient [8,9], and high-performance paths toward designated destinations [10]. However, given the direct implications for human life and traffic safety, there is increasing public and academic interest in accurately evaluating

the intelligence of autonomous path planning and comparing intelligence levels across different systems [11–21].

The current research landscape on quantitative evaluation of autonomous path planning intelligence is characterized by the following aspects. First, regarding the selection of evaluation metrics, many scholars typically employ performance metrics based on a system's internal parameters. For instance, Ref. [11] evaluates system performance using measurable parameters like trajectory tracking accuracy and computational real-time performance. Ref. [12] establishes a "white-box" evaluation model by incorporating control stability indices and sensor noise characteristics, while Ref. [13] employs dedicated metrics such as energy consumption models and motion smoothness to classify system intelligence levels. Although these methods offer precise insights into specific system performances, their effectiveness heavily relies on a comprehensive understanding of the internal architecture, assuming "white-box" access. This inherent dependency, however, severely limits their applicability for evaluating proprietary or architecturally complex systems where such internal knowledge is inaccessible. Second, concerning the selection of weighting methods, prevailing approaches largely utilize quantitative techniques that assume independence among metrics. Ref. [14] applies the Analytic Hierarchy Process (AHP) to construct judgment matrices through expert questionnaires. Ref. [15] employs the entropy weight method to calculate objective weights based on the dispersion degree of metric data, and Ref. [16] determines core indicator weights through factor analysis for dimensionality reduction. While these methods enhance the objectivity of evaluations, their foundational assumption of metric independence is often unrealistic. This neglect of inherent correlations can introduce bias into the weighting results, thereby compromising the validity of the final evaluation. Finally, addressing the combination of subjective and objective weights, researchers have proposed various combination weighting schemes. Ref. [17] proposed combining subjective AHP weights and objective entropy weights through linear weighting. Ref. [18] adopted a multiplicative synthesis rule to fuse weights derived from fuzzy AHP and the CRITIC method, and Ref. [19] balanced expert experience with data-driven results by setting fixed proportional coefficients. However, these methods lack a rigorous theoretical basis for determining the combination coefficients, often relying on heuristic or experiential settings. This reliance introduces subjectivity and arbitrariness into the process, ultimately undermining the robustness and scientific credibility of the final weights.

To overcome the limitations mentioned above, this paper proposes a Game-Theoretic Kendall's Coefficient (GTKC) weighting framework that addresses three critical challenges in autonomous path planning intelligence evaluation through a triple-pronged approach.

(1) We construct a three-dimensional evaluation metric system encompassing safety, efficiency, and comfort, explicitly modeling interdependencies among these metrics while ensuring observability and applicability across both "white-box" and "black-box".
(2) Since intelligence evaluation metrics are often interdependent, we accordingly select the improved Analytic Network Process (ANP) for subjective weighting due to its capacity to model metric correlations and feedback loops, and employ the CRITIC method for objective weighting as it effectively quantifies data contrast and conflict through standard deviation and correlation analysis.
(3) We introduce a game-theoretic optimization model that dynamically balances subjective and objective weights by minimizing deviations through Nash equilibrium, while rigorously evaluating internal consistency using Kendall's coefficient (for expert consensus) and coefficient of variation (for data stability). This framework harmonizes expert knowledge with data-driven insights, enhances robustness through credibility-weighted vector fusion, and significantly improves ranking consistency compared to conventional combination methods.

The remainder of this paper is structured as follows: Section 2 formally defines the problem of autonomous path planning intelligence evaluation. Section 3 presents the construction of the three-dimensional evaluation metric system and introduces the GTKC weighting framework. Section 4 validates GTKC against comparative methods through progressive experiments. Finally, Section 5 summarizes the key findings, highlights the main contributions of this research, and discusses potential avenues for future work.

2. Problem Description

To evaluate the intelligence of mobile applications, the research methodology should follow the evaluation framework illustrated in Figure 1.

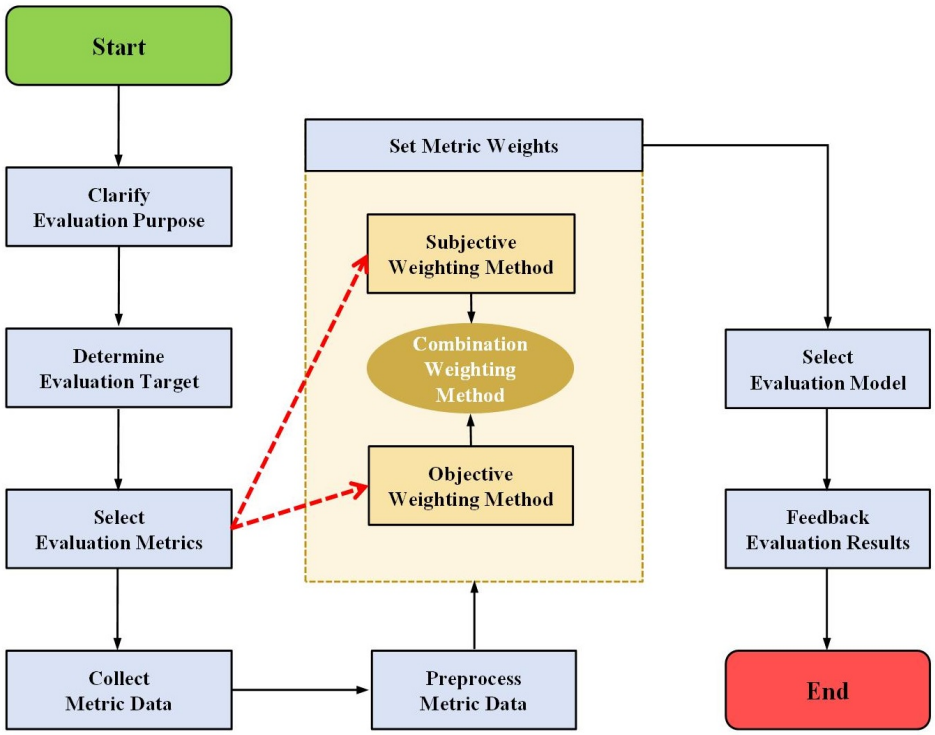

Figure 1. An illustration of the systematic evaluation framework flowchart. The solid arrow represents the process sequence, the red dashed arrow signifies a major influence on this process, and the yellow dashed box indicates the critical stage.

Let $A = a_1, a_2, \ldots, a_m$ be a set of m APPS algorithms under evaluation, $M = m_1, m_2, \ldots, m_n$ be n evaluation metrics, and $X \in \mathbb{R}^{m \times n}$ be the normalized performance matrix derived from observable behavioral data. The intelligence evaluation problem aims to construct a reliable ranking $\mathcal{R}$ of algorithms based on a weighted aggregation of metric scores:

$$\text{Score}(a_i) = \sum_{j=1}^{n} w_j \cdot X_{ij}, \quad \text{with} \quad \sum_{j=1}^{n} w_j = 1, \quad w_j \geq 0 \tag{1}$$

where $w = [w_1, w_2, \ldots, w_n]^T$ represents the metric weight vector, which quantifies the relative importance of each metric in the overall intelligence evaluation. The determination of an optimal weight vector w is crucial, as it directly influences the reliability and validity of the evaluation results.

Current approaches for determining w can be broadly categorized into three paradigms:

Subjective weighting methods derive weights based on expert knowledge and preferences:

$$w_s = f_s(E) \tag{2}$$

where E represents expert judgment matrices and f_s denotes subjective weighting functions such as Analytic Hierarchy Process (AHP) or Analytic Network Process (ANP).

Objective weighting methods compute weights directly from the data characteristics:

$$w_o = f_o(X) \tag{3}$$

where f_o represents objective weighting functions such as Entropy Weight Method (EWM) or Criteria Importance Through Intercriteria Correlation (CRITIC).

Combination weighting methods integrate both subjective and objective perspectives through linear combination:

$$w_c = \alpha w_s + \beta w_o, \quad \text{with} \quad \alpha + \beta = 1, \quad \alpha, \beta \geq 0 \tag{4}$$

For complex evaluation problems such as APPS intelligence evaluation, combination weighting approaches are generally preferred as they balance expert domain knowledge with empirical data patterns. However, the critical challenge lies in determining the optimal combination coefficients (α, β) that properly reflect the relative credibility of subjective and objective components. The selection of appropriate weighting methods and the determination of combination coefficients should be guided by the specific characteristics of the evaluation metrics and the inherent limitations of existing approaches.

3. Solution

3.1. Evaluation Metrics

We develop a hierarchical evaluation metric system for Autonomous Path Planning System (APPS) intelligence, structured around three core dimensions: safety, efficiency, and comfort. As illustrated in Figure 2, this system includes eight specific metrics, each designed to capture an essential aspect of path planning intelligence. Its design is grounded in the foundational principles of observability (metrics derivable from external outputs), quantifiability (numerically representable for objective comparison), and strong relevance (directly related to APPS operational objectives), and synthesizes key insights from existing literature [22–25].

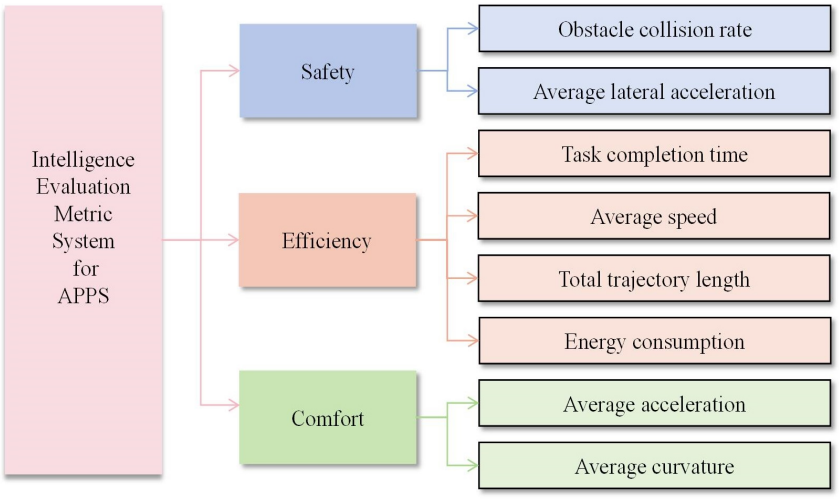

Figure 2. Multidimensional intelligence evaluation framework for APPS: quantifying system performance through safety, efficiency, and comfort metrics.

(1) Safety. This dimension quantifies the APPS fundamental capability to safeguard occupants, pedestrians, and surrounding vehicles during navigation, measured by:

Obstacle Collision Rate (C_r): The ratio of collision occurrences to total navigation attempts (dimensionless, expressed as a decimal):

$$C_r = \frac{N_c}{N_t} \tag{5}$$

where N_c is the number of collisions and N_t is the total number of navigation attempts.

Average Lateral Acceleration (A_l): The mean absolute value of lateral acceleration during navigation (m/s^2):

$$A_l = \frac{1}{T}\sum_{t=1}^{T}|a_l(t)| \tag{6}$$

where $a_l(t)$ is the lateral acceleration at time t and T is the total navigation duration.

(2) Efficiency. This dimension captures the performance of APPS in resource optimization and task execution timeliness, evaluated through:

Task Completion Time (T_c): The time required to complete the navigation task (s):

$$T_c = t_e - t_s \tag{7}$$

where t_s and t_e are the start and end times, respectively.

Average Speed (V_a): The mean velocity maintained during navigation (m/s):

$$V_a = \frac{1}{T}\int_0^T v(t)dt \tag{8}$$

Total Trajectory Length (L_t): The cumulative distance traveled along the actual path (m):

$$L_t = \sum_{k=1}^{N-1}\sqrt{(x_{k+1}-x_k)^2 + (y_{k+1}-y_k)^2} \tag{9}$$

where (x_k, y_k) represents the position coordinates at waypoint k.

Energy Consumption (E_c): The total energy expenditure during navigation (kWh):

$$E_c = \int_0^T P(t)dt \tag{10}$$

where $P(t)$ denotes the instantaneous power consumption at time t. For the purpose of comparing relative energy performance across algorithms under consistent vehicle dynamics, environment, and scenario, it can be simplified to the sum of squared accelerations $\int_0^T a^2(t)dt$ owing to the direct correlation between actuation power and acceleration.

(3) Comfort. This dimension reflects the quality of the passenger experience generated by the maneuvers of APPS, evaluated by:

Average Acceleration (A_a): The root mean square of acceleration magnitude (m/s^2):

$$A_a = \sqrt{\frac{1}{T}\int_0^T [a_x^2(t) + a_y^2(t)]dt} \tag{11}$$

Average Curvature (κ_a): The mean curvature of the traversed path (1/m):

$$\kappa_a = \frac{1}{L_t}\int_0^{L_t}|\kappa(s)|ds \tag{12}$$

where $\kappa(s)$ is the path curvature at arc length s.

3.2. Game-Theoretic Kendall's Coefficient Weighting Framework

In the process of combination weighting, the allocation of combination coefficients should be dynamically adjusted based on the internal consistency levels of both subjective and objective weights. Specifically, the higher the internal consistency of either subjective or objective weights, the larger the corresponding combination coefficient should be. Thus, the combination coefficients must be positively correlated with the internal consistency levels of the respective weights, so as to ensure that the more reliable weights exert greater influence in the combination weights.

Based on the above, the following section elaborates on the framework and implementation.

3.2.1. Framework

The GTKC weighting framework, schematically illustrated in Figure 3, is structured around three sequential stages to achieve a scientifically grounded fusion of subjective and objective perspectives.

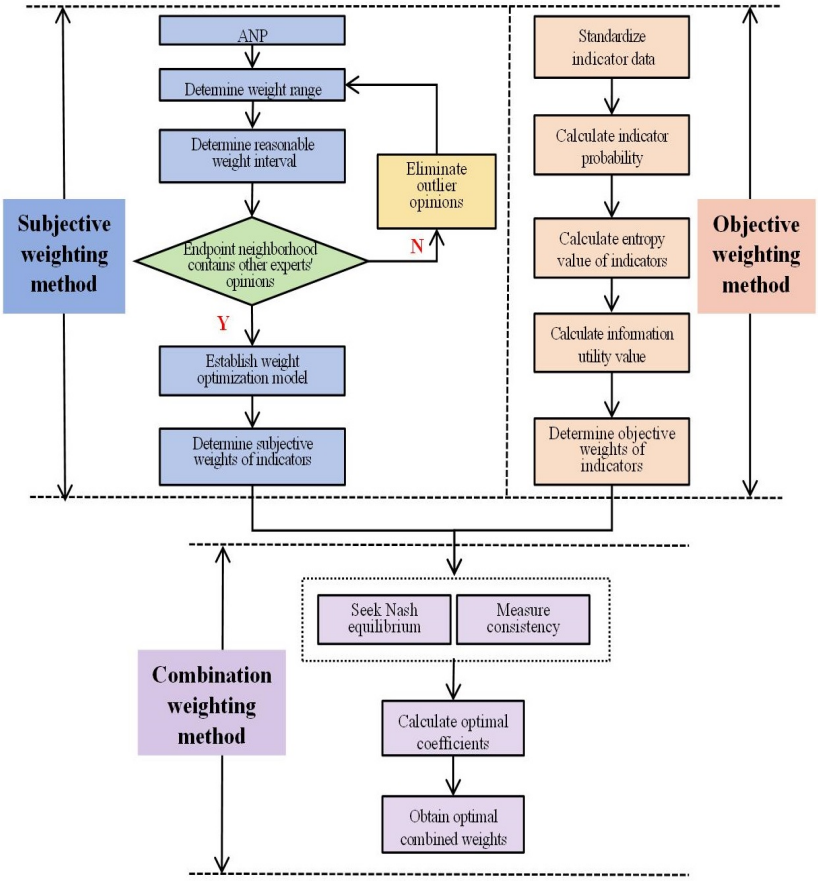

Figure 3. Implementation process of determining combination weights of evaluation metrics.

(1) Subjective Weighting Method: Expert judgments are aggregated and refined to derive subjective weights, explicitly accounting for interdependencies among evaluation metrics. This stage employs an enhanced group Analytic Network Process incorporating outlier filtering and optimization-based consensus fusion to enhance credibility.

(2) Objective Weighting Method: Data-driven weights are computed by quantifying both the inherent contrast intensity (dispersion) within each metric and the conflict (redundancy) between metrics based on their correlation structure.

(3) Combination Weighting Method: The subjective and objective weight vectors are optimally combined using a Nash equilibrium model minimizing deviation. Crucially, Kendall's coefficient and the coefficient of variation are introduced to dynami-

cally adjust the combination coefficients, assigning greater weight to the vector demonstrating higher internal consistency.

This tripartite framework facilitates the principled integration of expert knowledge and empirical data characteristics, dynamically balancing their influence based on inherent agreement.

3.2.2. Implementation

(1) Determining Subjective Weight Vector

ANP is well-suited for evaluating systems with internal feedback and complex interdependencies among metrics. Unlike AHP, which assumes metric independence, ANP enables a more realistic modeling of relationships, yielding more accurate results. However, when ANP is applied in group decision-making scenarios, expert judgments often exhibit inconsistencies. Simply discarding outlier opinions may lead to information loss and distorted evaluations.

To address this issue, this study adopts an improved group decision-making approach based on ANP. The implementation involves the following steps [26,27]:

(a) Determine Weight Ranges. Let the expert evaluation set be $F_m = \{f_{m1}, f_{m2}, \ldots, f_{mn}\}$, and let the weight set for metric c_j be $W_j = \{f_{1j}, f_{2j}, \ldots, f_{mj}\}$. The range of expert opinions is defined as

$$\omega_j^- = \min(f_{1j}, f_{2j}, \ldots, f_{mj}) \tag{13}$$

$$\omega_j^+ = \max(f_{1j}, f_{2j}, \ldots, f_{mj}) \tag{14}$$

(b) Identify Valid Intervals. Outliers are often located at the range boundaries. The interval length $d_j = \omega_j^+ - \omega_j^-$ reflects expert consensus. A large d_j implies strong disagreement. So, the singularity test uses $\delta = d_j/2$ the precision threshold.

(c) Eliminate Singular Points. The endpoint is considered an outlier and removed. This step is iteratively performed until all remaining values are within a valid interval.

(d) Construct Filtered Expert Matrix. Let m' be the number of experts remaining after outlier elimination. The filtered expert matrix is denoted as $F_{m'} = \{f_{m'1}, f_{m'2}, \ldots, f_{m'n}\}$.

(e) Optimize Subjective Weight. The final group-consensus weight ω_j^* is obtained by minimizing the deviation between expert opinions and the group weight:

$$\min \sum_{j=1}^{n} \sum_{m=1}^{m^*} \left(f_{mj} - \omega_j^*\right)^2 \tag{15}$$

Subject to:

$$\sum_{j=1}^{n} \omega_j^* = 1 \tag{16}$$

$$\omega_j^- < \omega_j^* < \omega_j^+. \tag{17}$$

The resulting subjective weight vector is

$$\omega_s = \left\{\omega_1^*, \omega_2^*, \ldots, \omega_j^*\right\}. \tag{18}$$

(2) Determining Objective Weight Vector

The CRITIC method is a well-established objective weighting technique that quantifies metric weights based on both data dispersion and inter-metric conflict. Its five-step procedure is as follows [28]:

(a) Data Normalization. To eliminate the influence of differing measurement scales, the raw data are normalized using min–max scaling.

For the cost-type metrics, the normalization formula is

$$x' = \frac{x_{\max} - x}{x_{\max} - x_{\min}} \tag{19}$$

For the benefit-type metric, the normalization formula is

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{20}$$

Thus, after normalization, a score of 1 always represents the best possible performance for any metric, and 0 the worst, ensuring a consistent and correct interpretation during the linear weighting in Equation (1).

(b) Calculate Standard Deviation s_j. The variability of each metric j is measured by its standard deviation S_j:

$$S_j = \sqrt{\sum_{i=1}^{n} \left(X'_{ij} - \overline{X_j}\right)^2 / (n-1)} \tag{21}$$

where X'_{ij} is the normalized value of the i-th sample for metric j, $\overline{X_j}$ is the mean, and n is the sample size.

(c) Calculate Conflict Between Metrics. The conflict or redundancy of metric j with respect to other metrics is given by

$$R_j = \sum_{i=1}^{p} (1 - r_{ij}) \tag{22}$$

where r_{ij} is the Pearson correlation coefficient between metric j and metric i, and p is the total number of metrics.

(d) Calculate Information Content. The amount of information carried by each metric is

$$C_j = S_j R_j \tag{23}$$

(e) Derive Objective Weight. The final objective weight for each metric is calculated as

$$\omega'_j = C_j / \sum_{j=1}^{p} C_j \tag{24}$$

This yields the objective weight vector:

$$\omega_o = \left\{\omega'_1, \omega'_2, \ldots, \omega'_j\right\}. \tag{25}$$

(3) Determining Combination Weight Vector

(a) Game-Theoretic Optimization Model

Let ω_1 and ω_2 denote the subjective weight and objective weight vectors, respectively. The combination weight vector is given by

$$\omega = \alpha_1 \omega_s + \alpha_2 \omega_o \tag{26}$$

The optimal values of α_1 and α_2 are determined by minimizing the total deviation from both components under a Nash equilibrium formulation:

$$\min f(\alpha_1, \alpha_2) = \left\|\omega - \omega_s\right\|_2^2 + \left\|\omega - \omega_o\right\|_2^2 \tag{27}$$

that is also subject to

$$\alpha_1 + \alpha_2 = 1, \alpha_1, \alpha_2 \in [0, 1] \tag{28}$$

The initial combination coefficients α_1 and α_2 are derived by solving the optimization problem defined in Equations (27) and (28). Substituting the linear combination $\omega = \alpha_1\omega_s + \alpha_2\omega_o$ and the constraint $\alpha_2 = 1 - \alpha_1$ into the objective function, we obtain

$$f(\alpha_1) = \|(\alpha_1\omega_s + (1-\alpha_1)\omega_o) - \omega_s\|_2^2 + \|(\alpha_1\omega_s + (1-\alpha_1)\omega_o) - \omega_o\|_2^2.$$

Simplifying the two terms,

$$f(\alpha_1) = \|(\alpha_1 - 1)(\omega_s - \omega_o)\|_2^2 + \|\alpha_1(\omega_s - \omega_o)\|_2^2 = (1-\alpha_1)^2\|\omega_s - \omega_o\|_2^2 + \alpha_1^2\|\omega_s - \omega_o\|_2^2.$$

Factoring out the common term,

$$f(\alpha_1) = \|\omega_s - \omega_o\|_2^2\left[(1-\alpha_1)^2 + \alpha_1^2\right] = \|\omega_s - \omega_o\|_2^2(2\alpha_1^2 - 2\alpha_1 + 1).$$

This is a univariate quadratic function in α_1. The first and second derivatives are

$$\frac{df}{d\alpha_1} = \|\omega_s - \omega_o\|_2^2(4\alpha_1 - 2), \quad \frac{d^2f}{d\alpha_1^2} = 4\|\omega_s - \omega_o\|_2^2.$$

Since $\frac{d^2f}{d\alpha_1^2} > 0$ for any $\omega_s \neq \omega_o$, the function $f(\alpha_1)$ is strictly convex and has a unique minimum. Setting the first derivative to zero,

$$\|\omega_s - \omega_o\|_2^2(4\alpha_1 - 2) = 0 \Rightarrow \alpha_1 = \frac{1}{2}.$$

Thus, the initial optimal coefficients are $\alpha_1' = \alpha_2' = 0.5$.

(b) Measurement Consistency

To comprehensively reflect the reliability of subjective and objective weights, different consistency measurement methods are adopted based on their inherent characteristics:

For subjective weights (derived from expert consensus), Kendall's coefficient of concordance is used to quantify the consistency of expert judgments [29,30]:

$$W_1 = 12S/m^2\left(n^3 - n\right) \tag{29}$$

where $S = \sum_{j=1}^{n}\left(X_j - \bar{X}\right)^2$; W_1 is the consistency levels of the subjective weights; X_j is the total rank score for metric j; $\bar{X}$ is the mean rank; m is the number of experts; n is the number of metrics.

For objective weights (derived from data-driven calculation), the coefficient of variation (CV, denoted as W_2) is used to measure consistency:

$$W_2 = \sigma_{\omega_2}/\mu_{\omega_2} \tag{30}$$

where σ_{ω_2} is the standard deviation of the objective weight vector ω_2; μ_{ω_2} is the mean of the objective weight vector ω_2.

Owing to the inherent "lower-is-better" nature of W_2, it is imperative to transform it into a "higher-is-better" measure normalized to the interval (0,1]. This transformation is critical for establishing syntactic comparability with W_1, as measures with aligned evaluation criteria (i.e., both "higher-is-better") and standardized scales are prerequisite for meaningful integration and subsequent analysis. The specific functional form employed to achieve this dual objective (semantic inversion and scale standardization) is explicitly defined as follows:

$$W_2' = 1/(1 + W_2) \tag{31}$$

(c) Optimal Combination Coefficients

Subsequently, the optimal combination coefficients are adjusted by assigning higher weights to the component with better consistency:

$$\alpha = W_1\alpha_1' / W_1\alpha_1' + W_2'\alpha_2' \tag{32}$$

$$\beta = W_2'\alpha_2' / W_1\alpha_1' + W_2'\alpha_2' \tag{33}$$

It is obvious that the adjusted combination coefficients satisfy $\alpha + \beta = 1$.

(d) Optimal Combination Weight Vector

The optimal combined weight vector is then computed as follows:

$$\omega^* = \alpha\omega_s + \beta\omega_o \tag{34}$$

This formulation ensures that the more consistent component receives proportionally greater influence in the final combination, enhancing the credibility and robustness of the overall evaluation.

Algorithm 1 summarizes the computational procedure of the proposed combination weighting framework based on GTKC.

Algorithm 1 GTKC

Require: Expert evaluation matrix $F_{s\tilde{w}}$, Objective index data X
Ensure: Optimal combination weights ω^*
1: **// Stage 1: Determine subjective weights ω_s via improved ANP**
2: Initialize empty weight set for each metric.
3: **for** each metric j **do**
4: Determine weight range $\left[\omega_j^-, \omega_j^+\right]$ using Equations (13) and (14)
5: Identify valid weight interval via singularity test
6: Remove outlier opinions and update expert set
7: Formulate filtered expert matrix $F_{s\tilde{w}}'$
8: **end for**
9: Solve optimization model via Equations (15)–(18) to get ω_s
10: **// Stage 2: Determine objective weights ω_o via CRITIC method**
11: Normalize objective data matrix X using Equations (19) and (20)
12: Calculate standard deviation S_j and correlation matrix for all metrics via Equation (21)
13: Compute conflict index R_j and information content C_j via Equations (22) and (23)
14: Derive ω_o by normalizing C_j across all metrics via Equations (24) and (25)
15: **// Stage 3: Determine optimal combination weights ω^***
16: Establish game model and compute initial α_1', α_2' via Equations (26)–(28)
17: Evaluate consistency: compute W_1 (expert consensus) via Equation (29) and W_2 (data variation) via Equation (30)
18: Transform W_2 to W_2' via Equation (31) to align with "higher-is-better" principle
19: Compute final coefficients α, β via credibility-weighted fusion in Equations (32) and (33)
20: Determine $\omega^* = \alpha\omega_s + \beta\omega_o$ via Equation (34)
21: **return** ω^*

4. Experiment

To systematically validate the proposed GTKC framework, this study designs three case scenarios with a progressive experimental logic: starting with verifying whether the framework can accurately evaluate intelligence (effectiveness), then moving to testing whether results are reliable for the same object (stability), and finally exploring whether it can stably distinguish between multiple objects (ranking consistency). This layered design ensures each case builds on the previous one, gradually deepening the verification of the framework's performance. Detailed experimental design and implementation are presented below:

4.1. Experimental Design

The primary goal of the designed testing scenario is not to evaluate the performance of the algorithms under evaluation, but to generate quantifiable behavioral data within the testing scenario for analyzing the effectiveness, stability, and ranking consistency of the proposed GTKC framework. If the testing scenario is excessively complex, the algorithms under evaluation will frequently fail—this will prevent the collection of sufficient successful trial data, making it impossible to conduct meaningful statistical comparisons of the various evaluation methods and ultimately undermining the aforementioned goal.

4.1.1. Test Scenarios

This experiment focuses on comparing the performance of multiple evaluation methods under identical conditions. To eliminate interference from environmental variables and road complexity, a 2D simulation environment was established to ensure consistent environmental conditions across all rounds, as illustrated in Figure 4.

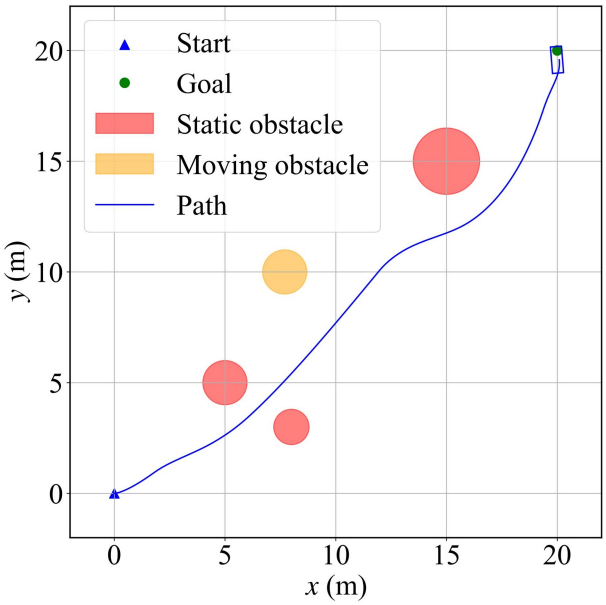

Figure 4. Standardized 2D simulation environment for evaluating APPS intelligence. The blue trajectory shows the planned path from the start (blue triangle) to the goal (green circle), successfully avoiding both static (red) and dynamic (yellow) obstacles.

(1) Simulation Environment Terrain Configuration: A flat rectangular area of 20 m × 20 m was constructed with a ground friction coefficient of 0.8, eliminating terrain irregularities that could affect kinematic modeling.

Obstacle: All obstacles were fully observable. Three static obstacles with radii of 1.5 m, 1 m, and 0.5 m were positioned at coordinates (15,15), (8,3), and (5,5), respectively. One dynamic obstacle (radius: 1 m) performed a linear reciprocating motion along the x-axis at a constant speed of 1m/s between (5,10) and (15,10).

Vehicle: The vehicle was modeled with dimensions of 1 m (length) × 0.5 m (width), a maximum acceleration of ± 1 m/s^2, and a differential drive system featuring a maximum steering angle of $\pm 30^\circ$ and a safety distance of 4m.

While deliberately simplified to ensure data collection viability across all algorithms under evaluation, this environment retains core challenges such as static/dynamic obstacle interactions, requisite path planning, and the need for timely decision-making. This ensures that the generated performance metrics are meaningful for intelligence evaluation while guaranteeing the reproducibility and generalizability of our comparative findings on evaluation methods.

(2) Task

The vehicle was required to navigate from the start point (0,0) to the endpoint (20,20) from a stationary state, prioritizing path planning while minimizing energy consumption and travel time.

(3) Experimental Protocol

Algorithms in Groups A and B underwent independent repeated trials under identical conditions. To maintain fixed strategies and consistency, the self-learning functionality of all RL algorithms was disabled, retaining only their decision execution capabilities.

To mitigate random fluctuations, testing was terminated when the variance of all metrics fell below a predefined threshold ϵ.

Five domain experts were invited for this experiment (two in intelligent vehicle engineering, one each in operations research and optimization, control theory, and human factor engineering) to provide multi-dimensional support for evaluation metric weight setting, covering simulation-reality adaptability, multi-objective optimization-based mathematical logic, simulation feasibility, and user experience consistency, thereby ensuring its scientificity and rationality.

4.1.2. Logic of Progressive Verification

To evaluate the performance of an evaluation method, a progressive verification of its results must be conducted through three critical aspects: effectiveness, stability, and ranking consistency [31]. The logic of progressive validation is outlined as follows:

(1) Case 1: Effectiveness. Addresses "Can the method accurately evaluate the intelligence level of APPS?" This is the basic prerequisite -only if the evaluation results are consistent with the actual intelligence level can subsequent reliability and discriminability verification be meaningful.

(2) Case 2: Stability. Addresses "Are the evaluation results of the same method stable across repeated trials?" This is a deepening of effectiveness—even if a method is accurate in a single trial, large fluctuations in repeated tests will reduce its practical value.

(3) Case 3: Ranking Consistency. Addresses "Can the method stably rank multiple APPS across repeated trials?" This is the application-oriented extension-APPS evaluation often requires comparing multiple systems, and stable ranking results are key to supporting decision-making.

4.1.3. Comparative Methods

To comprehensively verify the advantages of GTKC, three representative weighting methods covering different paradigms are selected as comparative methods. These methods are used in all three case scenarios of experiments to ensure consistent cross-method comparisons:

(1) AHP [14]: A purely subjective weighting approach based on expert judgment, representing traditional subjective evaluation (weak in objectivity and metric dependency).

(2) AHP-EWM [17]: A hybrid method combining AHP (subjective) and EWM (objective), representing early combination weighting (ignores metric interdependencies).

(3) ANP-CRITIC [19]: A correlation-aware combination method integrating ANP (subjective, handles dependencies) and CRITIC (objective, reflects data characteristics), representing improved combination weighting (lacks dynamic consistency-driven coefficient adjustment).

4.1.4. Algorithms under Evaluation

To meet the different objectives of the three case scenarios, two groups of algorithms under evaluation are designed, with clear division of labor:

(1) Group A: SAC Algorithm Variants with Multi-Level Stepwise Interventions

Used in Case 1 (effectiveness verification). By introducing controllable interventions (decision delay, Gaussian noise) to SAC, we construct algorithms with explicit intelligence differences (higher delay/noise → lower intelligence). This provides clear ground-truth rankings, enabling quantitative measurement of evaluation accuracy. For this purpose, two categories of algorithm variants were deployed in the simulation environment, with 600 experimental trials conducted. Experimental data revealed the following:

(a) Decision delay and task execution efficiency exhibited significant negative correlations. As shown in Figure 5, when system decision delay increased to 400 ms, task completion time surged approximately to 200 s (a 300% increase) This nonlinear response indicates rapid intelligence degradation beyond the 300 ms latency threshold.

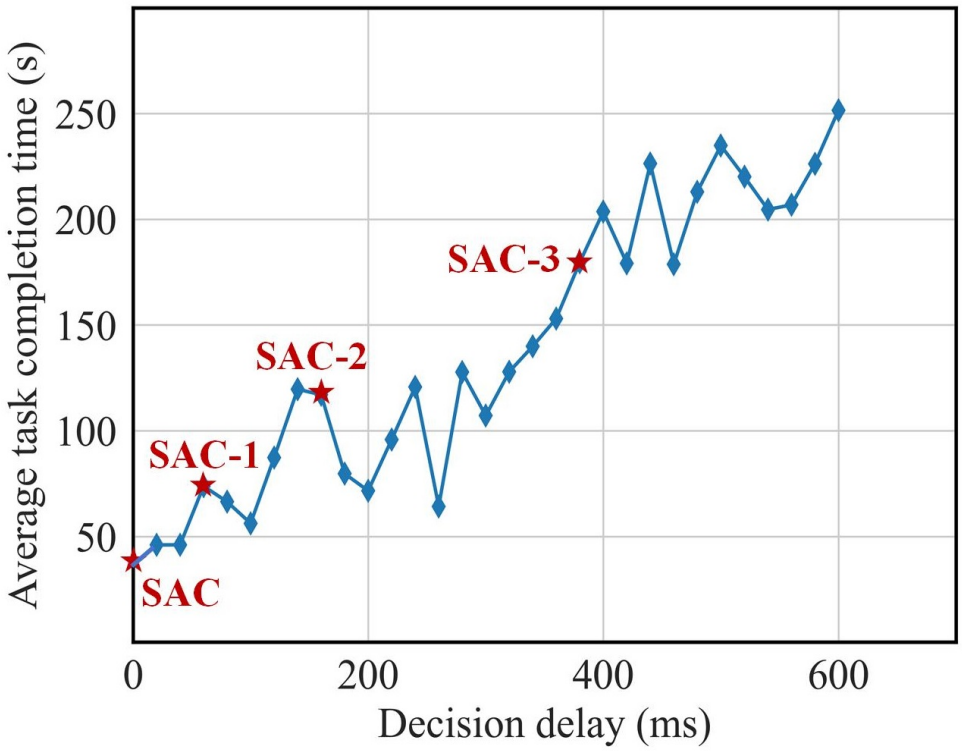

Figure 5. Decision delay vs. average task completion time. This figure presents the relationship between system decision delay (in milliseconds) and the resulting average task completion time (in seconds) observed during experimental trials on SAC algorithm variants.

(b) Gaussian noise intensity and algorithmic intelligence demonstrated marked negative correlations. This noise was injected into the control outputs of the algorithms to simulate actuation or control uncertainty. The injected noise was defined as zero-mean Gaussian white noise, with its standard deviation serving as the intervention magnitude (see Table 1). As depicted in Figure 6, when injected Gaussian noise standard deviation exceeded 0.2, collision rates escalated to 100% (a 100% decline relative to baseline conditions). This suggests that excessive noise (standard deviation > 0.1) destabilizes action outputs, triggering rapid intelligence deterioration.

Based on these findings, baseline algorithm variants exhibiting significant disparities in intelligence were selected from the experimental dataset to form Group A_1 and Group A_2. Their parameter configurations and reference rankings are detailed in Table 1. Specifically, the reference rankings provide a foundational basis for the subsequent comparative analysis of evaluation methods, thereby ensuring the reproducibility and comparability of the experimental conclusions.

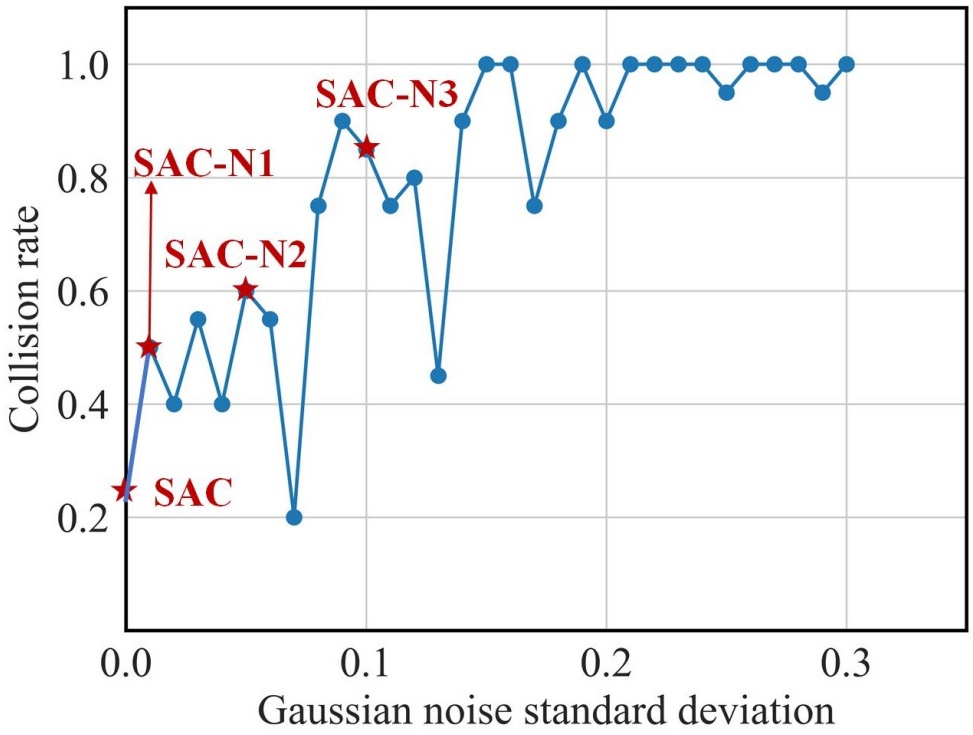

Figure 6. Gaussian noise standard deviation vs. collision rate. This figure illustrates the impact of injected Gaussian noise amplitude (standard deviation) on the collision rate (percentage) observed during experimental trials on SAC algorithm variants.

Table 1. Parameter configurations and reference intelligence rankings of Group A.

Group	Algorithm	Intervention Magnitude	Intelligence Ranking
A_1	SAC	0	1
	SAC-1	+60 ms	2
	SAC-2	+160 ms	3
	SAC-3	+380 ms	4
A_2	SAC	0	1
	SAC-N1	0.01	2
	SAC-N2	0.05	3
	SAC-N3	0.10	4

Group A_1: SAC algorithms with controlled decision delay interventions; Group A_2: SAC algorithms with controlled Gaussian noise amplitude interventions.

(2) Group B: 5 Representative Path Planning Algorithms

Used in Case 2 (stability verification) and Case 3 (ranking consistency verification). For Group B, algorithms under evaluation include five widely adopted path planning algorithms. These algorithms are categorized into two paradigms: Traditional Algorithms (DWA and Rapidly exploring Random Tree (RRT)) and RL Algorithms (A2C, Proximal Policy Optimization (PPO), and SAC).

DWA: Local path planning with real-time path planning but limited global optimization [32].

RRT: Sampling-based global planning suitable for high-dimensional spaces but with suboptimal trajectory smoothness [33].

A2C: Balances exploration and exploitation via policy gradients and value functions [34].

PPO: Enhances training stability through clipped policy updates [34].

SAC: A model-free RL algorithm using maximum entropy optimization to balance exploration and policy stability through reward-entropy maximization [35].

4.1.5. Experimental Procedure

All three case scenarios follow the same data collection and processing workflow to ensure result comparability:

Phase 1: Data Acquisition. For each algorithm, collect raw data of 8 metrics (covering safety: obstacle collision rate, average lateral acceleration; efficiency: task completion time, average speed, total trajectory length, energy consumption; comfort: average acceleration, average curvature across multiple trials.

Phase 2: Data Preprocessing. Normalize all metrics using min-max scaling to eliminate dimensional differences.

Phase 3: Weight Calculation. Compute weights for the four methods detailed in Section 4.1.3. AHP-EWM and ANP-CRITIC use empirical 0.5 combination coefficients [36–40]; GTKC uses dynamic coefficients described in Section 3.2.2.

Phase 4: Scoring and Ranking. Calculate overall intelligence scores via linear weighting of normalized metrics, and generate algorithm rankings based on scores.

4.2. Case 1: Effectiveness

This case aims to confirm whether the evaluation results of GTKC align with the actual intelligence level of APPS. For this purpose, Group A (SAC variants with clear ground-truth rankings) is used as the test subject, and we quantify the deviation between each method's evaluation results and the reference ranking.

4.2.1. Experimental Setup

(1) Test Object: Group A (A_1 : decision delay intervention, A_2: Gaussian noise intervention).

(2) Trial Design: Set the global precision threshold to $\varepsilon = 0.05$; conduct 13 consecutive trials for Group A_1 and 20 trials for Group A_2.

(3) Evaluation Metrics: Use Mean Absolute Error (MAE, measures deviation from reference ranking) and mean Spearman's correlation coefficient $\bar{\rho}$ (measures ranking similarity, range $[-1, 1]$, higher = more consistent) for quantitative analysis:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|o_{ij} - o_i\right| \tag{35}$$

where o_{ij} is the rank of the i-th algorithm under the j-th evaluation method, o_i is the reference rank assigned by human experts, and n is the number of algorithms.

$$\bar{\rho} = \frac{1}{m}\sum_{k=1}^{m}\left(1 - 6\sum_{i=1}^{n} d_{ik}^2 / n\left(n^2 - 1\right)\right) \tag{36}$$

where d_{ik} is the ranking difference between the evaluation ranking of the i-th algorithm in the k-th evaluation and the standard ranking, m is the number of experiments, and n is the number of algorithms.

4.2.2. Results and Analysis

Four comparative methods were applied to evaluate these algorithms, yielding the metric weights under each method. For key metrics (e.g., Task Completion Time T_c), GTKC's weight allocation is more in line with the actual impact of intelligence.

(1) Analysis of Group A_1 (Decision Delay Intervention)

The derived metric weights for Group A_1 are presented in Table 2. These metric weights are calculated using four weighting methods (AHP, AHP-EWM, ANP-CRITIC, and GTKC). Examining the weight allocation for T_c reveals significant differences across methods:

Table 2. Evaluation metric weights for Group A_1 under four comparative methods.

Method	Metric Weights							
	Obstacle Collision Rate	Average Lateral Acceleration	Task Completion Time	Average Speed	Total Trajectory Length	Energy Consumption	Average Acceleration	Average Curvature
AHP	0.0747	0.0890	0.0506	0.2967	0.2877	0.0926	0.0485	0.0601
AHP-EWM	0.2792	0.0631	0.0602	0.1813	0.2474	0.0659	0.0530	0.0499
ANP-CRITIC	0.0811	0.0889	0.0925	0.1216	0.1473	0.1212	0.1690	0.1785
GTKC	0.0975	0.1060	0.1047	0.1348	0.2074	0.1349	0.1004	0.1143
Δ Weight (%)	20.22%	19.24%	13.19%	10.86%	40.80%	11.30%	−40.59%	−35.97%

Δ Weight (%), is calculated as $[W_{\text{GTKC}} - W_{\text{ANP-CRITIC}})/W_{\text{ANP-CRITIC}}] \times 100\%$.

- Comparison between AHP-EWM (0.0602) and ANP-CRITIC (0.0925) indicates that accounting for inter-metric relationships substantially influences weight allocation, leading to a 53.65% increase in weight for T_c.
- Further comparison between ANP-CRITIC (0.0925) and GTKC (0.1047) shows that the GTKC weight exceeds the ANP-CRITIC weight by 13.19%. This difference arises from the GTKC framework's dynamic determination of combination coefficients (subjective-to-objective ratio = 1.27), prioritizing the more credible subjective weights derived via the improved ANP process.

After normalizing the experimental data, linear weighting was applied to derive evaluation scores and corresponding rankings for each algorithm under the different evaluation methods. The rank order comparison against the reference ranking (Table 1) is illustrated in Figure 7. The dashed lines represent the ground-truth reference rankings.

(2) Analysis of Group A_2 (Gaussian Noise Intervention)

The derived metric weights for Group A_2 are presented in Table 3. These metric weights are calculated using four weighting methods (AHP, AHP-EWM, ANP-CRITIC, and GTKC). Similar to Group A_1, the weight allocation for T_c highlights methodological differences:

- The ANP-CRITIC weight (0.1061) is 32.63% higher than the AHP-EWM weight (0.08), again underscoring the impact of considering metric interdependencies.
- The GTKC weight (0.1289) surpasses the ANP-CRITIC weight (0.1061) by 21.49%. This increase is attributed to GTKC's consistency-driven adjustment (subjective-to-objective ratio = 1.29), favoring the more reliable subjective weights.

The evaluation scores and rankings derived for Group A_2 across the 20 trials are compared against the reference ranking in Figure 8.

Table 3. Evaluation metric weights for Group A_2 under four comparative methods.

Method	Metric Weights							
	Obstacle Collision Rate	Average Lateral Acceleration	Task Completion Time	Average Speed	Total Trajectory Length	Energy Consumption	Average Acceleration	Average Curvature
AHP	0.0747	0.0890	0.0506	0.2967	0.2877	0.0926	0.0485	0.0601
AHP-EWM	0.1682	0.0593	0.0800	0.2029	0.2958	0.0664	0.0728	0.0546
ANP-CRITIC	0.1033	0.0766	0.1061	0.1306	0.1356	0.1055	0.1751	0.1673
GTKC	0.1211	0.0868	0.1289	0.1424	0.1872	0.1204	0.1108	0.1025
Δ Weight (%)	17.23%	13.32%	21.49%	9.04%	38.05%	14.12%	−36.72%	−38.73%

Δ Weight (%), is calculated as $[W_{\text{GTKC}} - W_{\text{ANP-CRITIC}})/W_{\text{ANP-CRITIC}}] \times 100\%$.

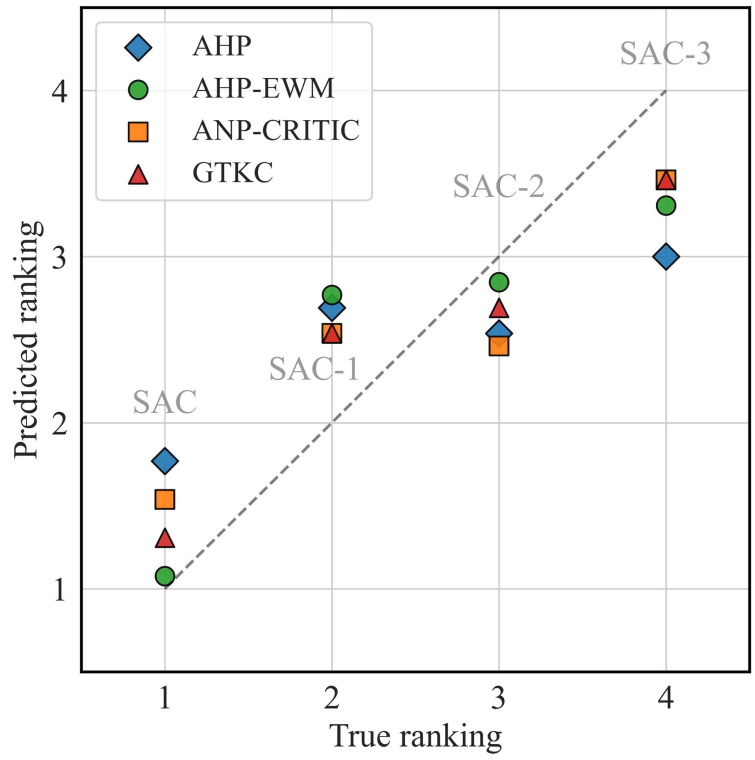

Figure 7. Comparative ranking performance of Group A_1. This figure compares the intelligence ranking outcomes of Group A_1 generated by four comparative methods. Each method's ranking deviation is visually quantified by vertical distance to the dashed reference lines.

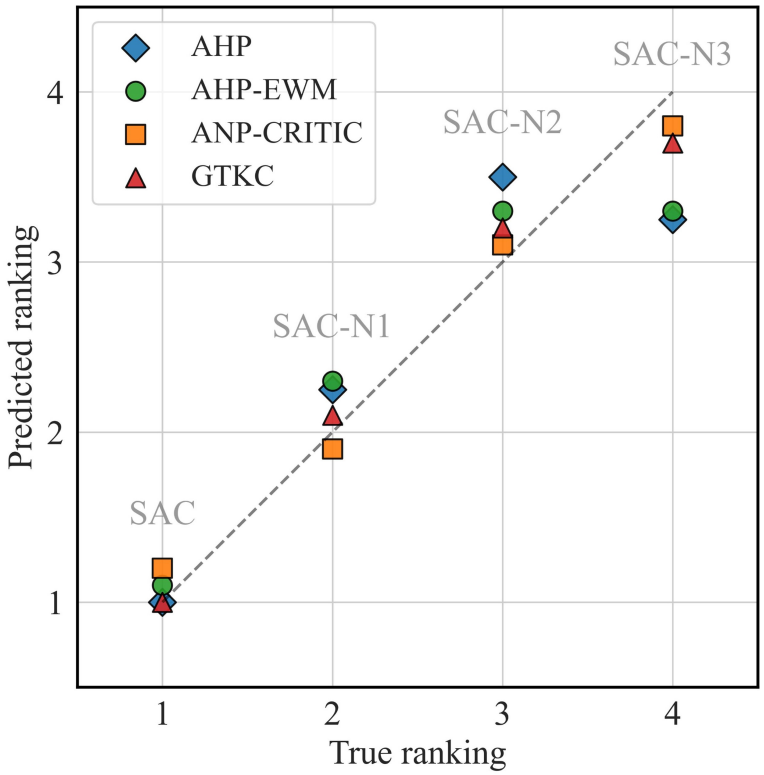

Figure 8. Comparative ranking performance of Group A_2. This figure compares the intelligence ranking outcomes of Group A_2 generated by four comparative methods. Each method's ranking deviation is visually quantified by vertical distance to the dashed reference lines.

MAE and $\overline{\rho}$ of each evaluation method are derived from Equations (35) and (36), as shown in Table 4. Consequently, across both algorithm groups subjected to distinct intelligence-diminishing interventions (decision delay and Gaussian noise), the GTKC framework consistently demonstrates superior effectiveness in approximating the ground-truth intelligence rankings, as evidenced by its significantly lower MAE and higher $\overline{\rho}$ compared to the comparative methods (AHP, AHP-EWM, ANP-CRITIC).

Table 4. Performance comparison of evaluation methods based on MAE and $\overline{\rho}$.

		AHP	AHP-EWM	ANP-CRITIC	GTKC
A_1	**SAC**	1.77	1.08	1.54	1.31
	SAC-1	2.69	2.77	2.54	2.54
	SAC-2	2.54	2.85	2.46	2.69
	SAC-3	3.00	3.31	3.46	3.46
	MAE	0.73	0.42	0.54	0.42
	$\overline{\rho}$	0.80	1.00	0.80	1.00
A_2	**SAC**	1.00	1.10	1.20	1.00
	SAC-N1	2.25	2.30	1.90	2.10
	SAC-N2	3.50	3.30	3.10	3.20
	SAC-N3	3.25	3.30	3.80	3.70
	MAE	0.38	0.35	0.15	0.15
	$\overline{\rho}$	0.80	0.95	1.00	1.00

MAE: Mean Absolute Error; $\overline{\rho}$: mean Spearman's correlation coefficient.

This confirms that GTKC can accurately evaluate APPS intelligence, laying the foundation for Case 2.

4.3. Case 2: Stability

Following the confirmation of effectiveness, this case aims to verify stability—specifically, whether the evaluation results of the same algorithm remain stable across repeated trials. We use Group B (five representative algorithms) as the test object, and measure the fluctuation of evaluation values to evaluate the method's reliability.

4.3.1. Experimental Setup

(1) Test Object: Group B (DWA, RRT, A2C, PPO, SAC).

(2) Trial Design: Conduct 17 consecutive independent trials; disable the self-learning function of RL algorithms to ensure fixed decision strategies.

(3) Evaluation Metrics: Use the mean of the Standard Deviations (MSD, the standard deviations of each algorithm's evaluation values across trials-lower MSD = better stability):

$$MSD = \frac{1}{n}\sum_{i=1}^{n}\sigma_i \tag{37}$$

where σ_i is the standard deviation of the i-th evaluation values, and n is the number of algorithms.

4.3.2. Results and Analysis

The resulting metric weights for each method are summarized in Table 5. The influence of considering inter-metric relationships on weight allocation is evident. For instance, comparing the weight assigned to T_c by AHP-EWM (0.0329) and ANP-CRITIC (0.063) reveals a significant 91.49% increase under the latter method. Further comparison between ANP-CRITIC (0.063) and GTKC (0.0787) for the same metric indicates that the dynamic adjustment based on subjective–objective consistency (with a coordination coefficient ratio

of 1.29 favoring the more credible subjective weights) resulted in GTKC assigning a weight 24.92% higher than ANP-CRITIC.

Subsequently, intelligence evaluation scores were computed for all Group B under each method. The distribution of these scores across the seventeen trials is presented in Figure 9:

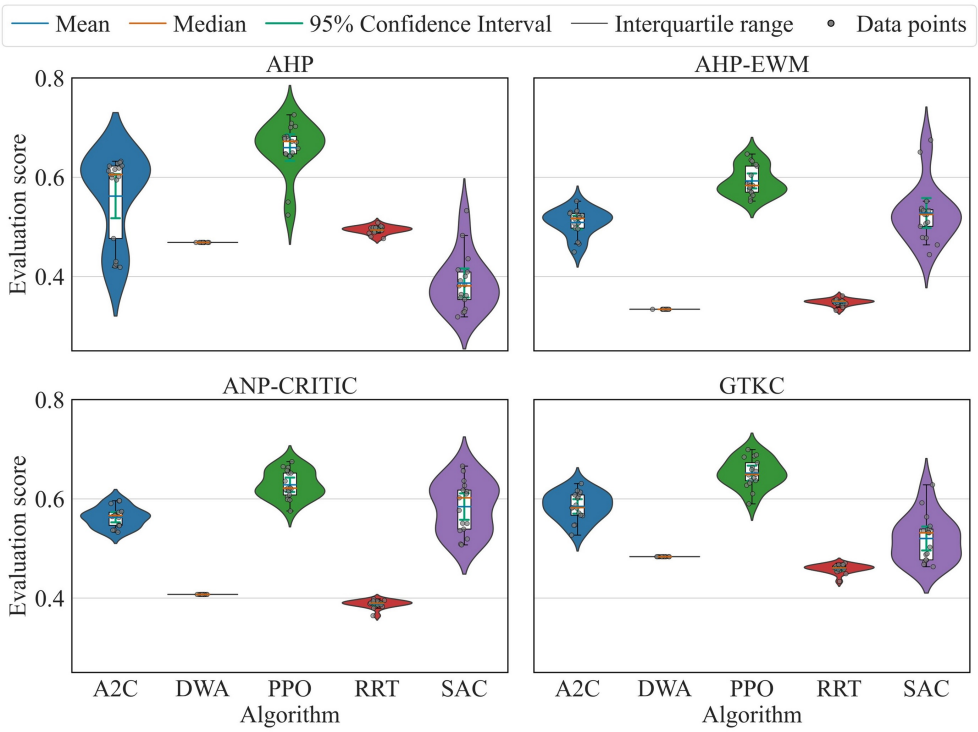

Figure 9. Distribution of intelligence evaluation scores for Group B. Violin plots depict the probability density and dispersion of scores across 17 independent trials, evaluated under four comparative methods for Group B.

- AHP: Algorithm scores demonstrated clear stratification but exhibited significant fluctuations, particularly for A2C and PPO.
- AHP-EWM: Evaluation values showed substantial discrepancies between algorithms, with SAC displaying pronounced upward score volatility.
- ANP-CRITIC: Moderate differences were observed between algorithm scores, and A2C showed considerably reduced fluctuations compared to AHP.
- GTKC: Algorithm scores were relatively well-differentiated and exhibited patterns broadly similar to those under ANP-CRITIC, suggesting comparable stability.

Table 5. Evaluation metric weights for Group B under four comparative methods.

Method	Metric Weights							
	Obstacle Collision Rate	**Average Lateral Acceleration**	**Task Completion Time**	**Average Speed**	**Total Trajectory Length**	**Energy Consumption**	**Average Acceleration**	**Average Curvature**
AHP	0.0747	0.0890	0.0506	0.2967	0.2877	0.0926	0.0485	0.0601
AHP-EWM	0.0764	0.0676	0.0329	0.1876	0.1980	0.0531	0.1825	0.2021
ANP-CRITIC	0.1290	0.0940	0.0630	0.1199	0.1049	0.0872	0.1993	0.2026
GTKC	0.1394	0.1107	0.0787	0.1334	0.1706	0.1051	0.1268	0.1353
Δ Weight (%)	8.06%	17.77%	24.92%	11.26%	62.63%	20.53%	−36.38%	−33.21%

Δ Weight (%), is calculated as $[W_{\text{GTKC}} - W_{\text{ANP-CRITIC}})/W_{\text{ANP-CRITIC}}] \times 100\%$.

The MSD of the evaluation values, calculated using Equation (37), provides a quantitative measure of stability across methods. Figure 10 visualizes these results. AHP exhibited the highest MSD, indicating the poorest stability. Although GTKC (MSD = 0.023) demonstrated a slightly higher MSD than ANP-CRITIC (MSD = 0.022), a paired t-test revealed that this difference was not statistically significant ($t(4) = 0.477$, $p = 0.658 > 0.05$). The 95% confidence interval for the difference was $[-0.005, 0.007]$, which includes zero, further confirming no significant difference in stability between the two methods. Both GTKC and ANP-CRITIC significantly outperformed AHP-EWM and AHP. Consequently, the stability ranking of the evaluation methods, ordered from most to least stable, is ANP-CRITIC $\approx$ GTKC $\succ$ AHP-EWM $\succ$ AHP.

This indicates that GTKC has good stability—its results for the same algorithm are reliable, providing support for the third case scenario of ranking consistency verification.

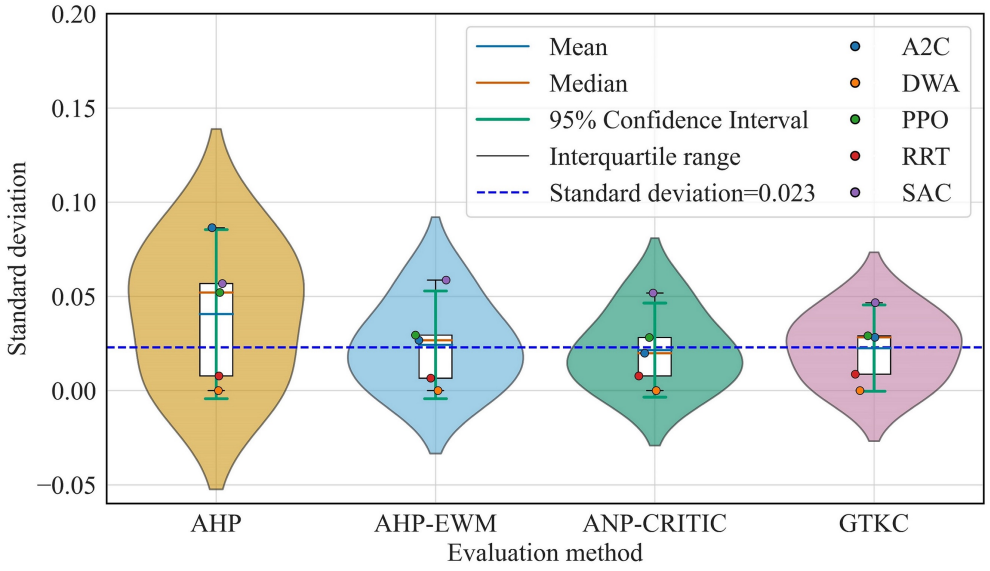

Figure 10. Stability comparison of comparative methods via standard deviation. This figure compares the stability characteristics of four comparative methods through vertical bar charts depicting the standard deviation values for Group B.

4.4. Case 3: Ranking Consistency

Building on Case 2, this case aims to verify ranking consistency—specifically, whether the ranking order of multiple algorithms remains consistent across repeated trials. Leveraging the 17 trials of Group B (from Case 2), we use Kendall's coefficient to quantify the agreement of ranking results, which reflects the method's discriminative ability for multiple algorithms.

4.4.1. Experimental Setup

(1) Test Object: Group B (DWA, RRT, A2C, PPO, SAC).

(2) Trial Design: Conduct 17 consecutive independent trials (from Section 4.3) to ensure consistency with stability verification.

(3) Evaluation Metrics: Use Kendall's coefficient W (range $[0, 1]$, higher W = more consistent ranking across trials):

$$W = 12S / M^2 \left(N^3 - N \right) \tag{38}$$

where S is the sum of squared deviations from the mean rank, M is the number of trials, and N is the number of algorithms.

4.4.2. Results and Analysis

As defined in Equation (38) and illustrated in Figure 11, the GTKC framework achieved the highest concordance ($W = 0.924$), indicating superior stability in the produced ranking orders compared to the alternative methods. The AHP-EWM combination method yielded a slightly lower ($W = 0.910$). In contrast, the purely subjective AHP approach demonstrated significantly lower reliability ($W = 0.812$), consistent with its inherent susceptibility to evaluator variability. The ANP-CRITIC method ($W = 0.892$) also exhibited lower concordance than GTKC and AHP-EWM. Consequently, the ranking consistency performance of the evaluated methods is ordered as GTKC $\succ$ AHP-EWM $\succ$ ANP-CRITIC $\succ$ AHP.

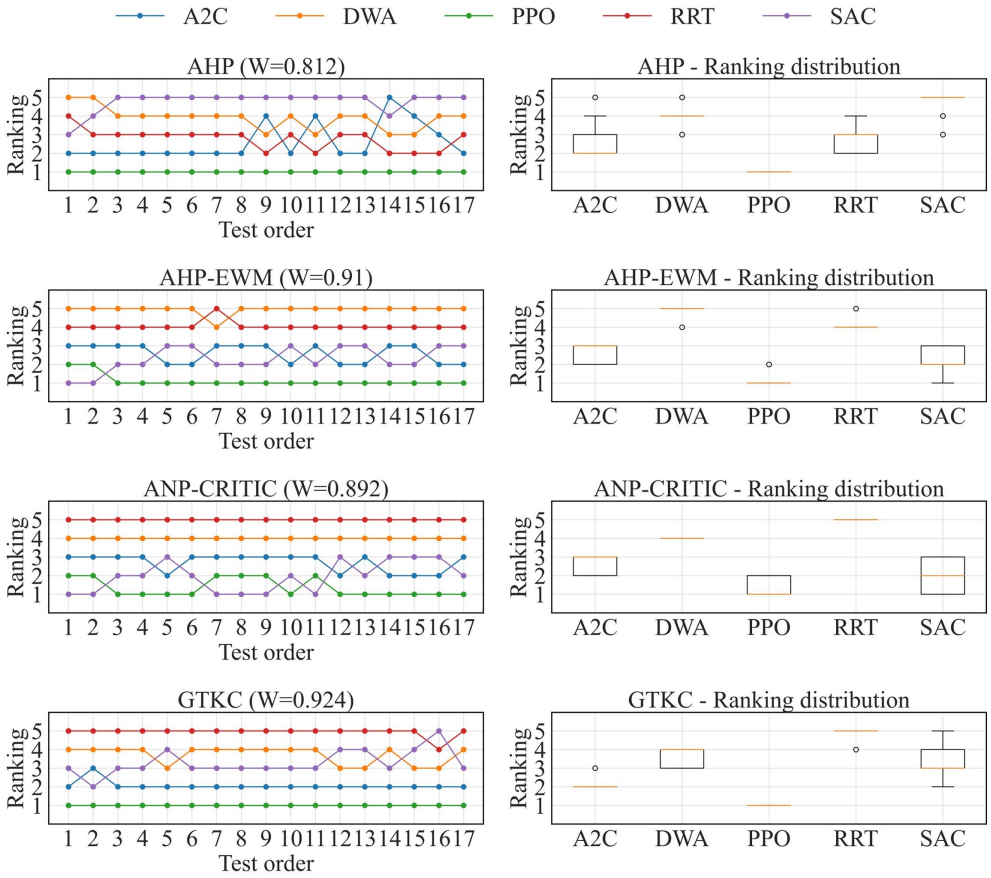

Figure 11. Ranking consistency analysis via Kendall's coefficient. This figure quantifies the ranking consistency of Group B across four comparative methods through 17 experimental trials.

This confirms that GTKC can stably rank multiple APPS, making it suitable for practical scenarios where multiple systems need to be compared.

5. Conclusions

This study proposes the Game-Theoretic Kendall's Coefficient (GTKC) weighting framework to address three core limitations in autonomous path planning intelligence evaluation: the lack of architecture-agnostic metrics, neglect of metric interdependencies, and non-rigorous weight fusion. Our solution integrates a three-dimensional metric system (Safety, Efficiency, Comfort) with improved ANP for subjective weights and CRITIC for objective weights, dynamically fused via a Game-Theoretic Kendall's Coefficient that prioritizes high-consistency weight vector.

Experimental validation demonstrates GTKC's superiority over comparative methods (AHP, AHP-EWM, ANP-CRITIC). In effectiveness tests, GTKC achieved the lowest *MAE* (0.15 for Group A_2) and perfect Spearman's correlation coefficient ($\overline{\rho} = 1.0$) with ground-

truth rankings, correctly classifying SAC variants (Table 4). While ANP-CRITIC showed a marginally lower *MSD* (0.022) than GTKC (0.023) in Group B, a paired *t*-test confirmed this difference was not statistically significant ($t(4) = 0.477$, $p = 0.658 > 0.05$), indicating comparable stability between the two methods. GTKC exhibited superior ranking consistency ($W = 0.924$) by dynamically balancing expert knowledge and data patterns (Figure 11). The framework advances APPS evaluation theory through its rigorous weight fusion mechanism and generalizable paradigm applicable to both traditional (DWA, RRT) and RL-based (A2C, PPO, SAC) algorithms.

This study has limitations including the use of a 2D simulation environment, insufficient real-world data validation, and the absence of statistical significance testing. To address these limitations, future work will extend experiments to real-world scenarios (e.g., sensor noise, delayed execution), validate method robustness through ablation studies, and employ statistical tests to support the generalizability of conclusions.

Author Contributions: Conceptualization, Z.D. and J.Y.; methodology, Z.D. and R.Y.; software, Z.D. and G.S.; validation, Z.D., J.Y. and M.L.; formal analysis, G.S.; investigation, J.Y. and M.L.; data curation, Z.D., J.Y. and M.L.; writing—original draft preparation, Z.D. and R.Y.; writing—review and editing, Z.D. and R.Y.; 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.

References

1. Abed, E.H.; Fu, J.H. Local feedback stabilization and bifurcation control, I. Hopf bifurcation. *Syst. Control Lett.* **1986**, *7*, 11–17. [CrossRef]
2. Li, S. Online Iterative Learning Enhanced Sim-to-Real Transfer for Efficient Manipulation of Deformable Objects. In Proceedings of the 2025 IEEE 19th International Symposium on Applied Computational Intelligence and Informatics (SACI), Timisoara, Romania, 19–24 May 2025; pp. 000015–000016.
3. Tan, M.; Ren, Y.; Pan, R.; Wang, L.; Chen, J. Fair and efficient electric vehicle charging scheduling optimization considering the maximum individual waiting time and operating cost. *IEEE Trans. Veh. Technol.* **2023**, *72*, 9808–9820. [CrossRef]
4. Cheng, S.; Wang, Z.; Yang, J.; Sun, Y.; Zhang, M.; Hou, S.; Lu, H.; Luo, C.; Shi, Y. Robot Path Planning in Unknown Environments based on a Learning-guided Optimization Approach. *Tsinghua Sci. Technol.* **2025**, forthcoming. [CrossRef]
5. Ge, J.; Zhang, J.; Zhang, Y.; Yao, D.; Zhang, Z.; Zhou, R. Autonomous Vehicles Testing Considering Utility-Based Operable Tasks. *Tsinghua Sci. Technol.* **2023**, *28*, 965–975. [CrossRef]
6. Abdallaoui, S.; Aglzim, E.H.; Chaibet, A.; Kribèche, A. Thorough review analysis of safe control of autonomous vehicles: Path planning and navigation techniques. *Energies* **2022**, *15*, 1358. [CrossRef]
7. Sánchez-Ibáñez, J.R.; Pérez-del Pulgar, C.J.; García-Cerezo, A. Path planning for autonomous mobile robots: A review. *Sensors* **2021**, *21*, 7898. [CrossRef]
8. Wang, H.; Lu, B.; Li, J.; Liu, T.; Xing, Y.; Lv, C.; Cao, D.; Li, J.; Zhang, J.; Hashemi, E. Risk assessment and mitigation in local path planning for autonomous vehicles with LSTM based predictive model. *IEEE Trans. Autom. Sci. Eng.* **2021**, *19*, 2738–2749. [CrossRef]
9. Lin, Z.; Wu, K.; Shen, R.; Yu, X.; Huang, S. An efficient and accurate A-star algorithm for autonomous vehicle path planning. *IEEE Trans. Veh. Technol.* **2023**, *73*, 9003–9008. [CrossRef]
10. Nawaz, M.; Tang, J.K.T.; Bibi, K.; Xiao, S.; Ho, H.P.; Yuan, W. Robust cognitive capability in autonomous driving using sensor fusion techniques: A survey. *IEEE Trans. Intell. Transp. Syst.* **2023**, *25*, 3228–3243. [CrossRef]
11. Zuo, Y.; Yang, C.; Li, S.; Wang, W.; Xiang, C.; Qie, T. A model predictive trajectory tracking control strategy for heavy-duty unmanned tracked vehicle using deep Koopman operator. *Eng. Appl. Artif. Intell.* **2025**, *159*, 111698. [CrossRef]

12. Ramírez, K.G.; Nuevo-Gallardo, C.; Ulecia, J.M.S.; Pozas, B.M.; Bandera, C.F. Digital Twin Implementation Based on a White-Box Building Energy Model: A Case Study on Blind Control for Passive Heating. *Energy Build.* **2025**, *349*, 116454. [CrossRef]
13. Tian, S.; Wei, C.; Jian, S.; Ji, Z. Preference-based deep reinforcement learning with automatic curriculum learning for map-free UGV navigation in factory-like environments. *Eng. Sci. Technol. Int. J.* **2025**, *70*, 102147. [CrossRef]
14. Sun, Y.; Xiong, G.; Song, W.; Gong, J.; Chen, H. Test and evaluation of autonomous ground vehicles. *Adv. Mech. Eng.* **2014**, *6*, 681326. [CrossRef]
15. Vilone, G.; Longo, L. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Inf. Fusion* **2021**, *76*, 89–106. [CrossRef]
16. Zhao, Y.N.; Meng, K.W.; Gao, L. The Entropy-Cost Function Evaluation Method for Unmanned Ground Vehicles. *Math. Probl. Eng.* **2015**, *2015*, 410796. [CrossRef]
17. Adhikari, R.; Agrawal, R. Performance evaluation of weights selection schemes for linear combination of multiple forecasts. *Artif. Intell. Rev.* **2014**, *42*, 529–548. [CrossRef]
18. Zhang, H.; Bai, X.; Hong, X. Site selection of nursing homes based on interval type-2 fuzzy AHP, CRITIC and improved TOPSIS methods. *J. Intell. Fuzzy Syst.* **2022**, *42*, 3789–3804. [CrossRef]
19. Chen, C.H. A novel multi-criteria decision-making model for building material supplier selection based on entropy-AHP weighted TOPSIS. *Entropy* **2020**, *22*, 259. [CrossRef] [PubMed]
20. Wang, C.; Qiao, J.; Huang, X.; Song, S.; Hou, H.; Jiang, T.; Rui, L.; Wang, J.; Sun, J. Apache IoTDB: A Time Series Database for Large Scale IoT Applications. *ACM Trans. Database Syst.* **2025**, *50*, 1–45. [CrossRef]
21. Liu, Y.; Zhang, J.; Chen, Y.; Wang, W.; Yang, S.; Na, X.; Sun, Y.; He, Y. Real-time continuous activity recognition with a commercial mmWave radar. *IEEE Trans. Mob. Comput.* **2024**, *24*, 1684–1698. [CrossRef]
22. Huang, H.; Zheng, X.; Yang, Y.; Liu, J.; Liu, W.; Wang, J. An integrated architecture for intelligence evaluation of automated vehicles. *Accid. Anal. Prev.* **2020**, *145*, 105681. [CrossRef] [PubMed]
23. Yan, Y.; Wen, H.; Deng, Y.; Chow, A.H.; Wu, Q.; Kuo, Y.H. A mixed-integer programming-based Q-learning approach for electric bus scheduling with multiple termini and service routes. *Transp. Res. Part C Emerg. Technol.* **2024**, *162*, 104570. [CrossRef]
24. Yang, J.; Wang, Z.; Wang, D.; Zhang, Y.; Lu, Q.; Feng, S. Adaptive safety performance testing for autonomous vehicles with adaptive importance sampling. *Transp. Res. Part C Emerg. Technol.* **2025**, *179*, 105256. [CrossRef]
25. Zhou, J.; Wang, L.; Wang, X. Scalable evaluation methods for autonomous vehicles. *Expert Syst. Appl.* **2024**, *249*, 123603. [CrossRef]
26. Agnusdei, L.; Krstić, M.; Palmi, P.; Miglietta, P.P. Digitalization as driver to achieve circularity in the agroindustry: A SWOT-ANP-ADAM approach. *Sci. Total Environ.* **2023**, *882*, 163441. [CrossRef]
27. Cui, C.; Li, B.; Chen, X. Group Decision-Making Method of Entry Policy During a Pandemic. *Tsinghua Sci. Technol.* **2024**, *29*, 56–65. [CrossRef]
28. Ilham, N.I.; Dahlan, N.Y.; Hussin, M.Z. Optimizing solar PV investments: A comprehensive decision-making index using CRITIC and TOPSIS. *Renew. Energy Focus* **2024**, *49*, 100551. [CrossRef]
29. Fritschy, C.; Spinler, S. The impact of autonomous trucks on business models in the automotive and logistics industry–a Delphi-based scenario study. *Technol. Forecast. Soc. Change* **2019**, *148*, 119736. [CrossRef]
30. van den Heuvel, E.; Zhan, Z. Myths about linear and monotonic associations: Pearson'sr, Spearman's ρ, and Kendall's τ. *Am. Stat.* **2022**, *76*, 44–52. [CrossRef]
31. Lee, S.; Lee, J.; Moon, H.; Park, C.; Seo, J.; Eo, S.; Koo, S.; Lim, H. A survey on evaluation metrics for machine translation. *Mathematics* **2023**, *11*, 1006. [CrossRef]
32. Dobrevski, M.; Skočaj, D. Dynamic adaptive dynamic window approach. *IEEE Trans. Robot.* **2024**, *40*, 3068–3081. [CrossRef]
33. Saleh, I.; Borhan, N.; Rahiman, W. Smoothing RRT Path for Mobile Robot Navigation Using Bioinspired Optimization Method. *Pertanika J. Sci. Technol.* **2024**, *32*, 2327–2342. [CrossRef]
34. Zhu, K.; Zhang, T. Deep reinforcement learning based mobile robot navigation: A review. *Tsinghua Sci. Technol.* **2021**, *26*, 674–691. [CrossRef]
35. Ding, F.; Ma, G.; Chen, Z.; Gao, J.; Li, P. Averaged Soft Actor-Critic for Deep Reinforcement Learning. *Complexity* **2021**, *2021*, 6658724. [CrossRef]
36. Karlova-Sergieva, V. Approach for the Assessment of Stability and Performance in the s-and z-Complex Domains. *Automation* **2025**, *6*, 61. [CrossRef]
37. Chen, J.; Gao, C.; Zhou, H.; Wang, Q.; She, L.; Qing, D.; Cao, C. Urban flood risk assessment based on a combination of subjective and objective multi-weight methods. *Appl. Sci.* **2024**, *14*, 3694. [CrossRef]
38. Pang, N.; Luo, W.; Wu, R.; Lan, H.; Qin, Y.; Su, Q. Safety evaluation of commercial vehicle driving behavior using the AHP—CRITIC algorithm. *J. Shanghai Jiaotong Univ. (Sci.)* **2023**, *28*, 126–135. [CrossRef]

39. Chang, K.H. Integrating subjective–objective weights consideration and a combined compromise solution method for handling supplier selection issues. *Systems* **2023**, *11*, 74. [CrossRef]
40. Dou, F.; Xing, H.; Li, X.; Yuan, F.; Lu, Z.; Li, X.; Ge, W. 3D geological suitability evaluation for urban underground space development based on combined weighting and improved TOPSIS. *Nat. Resour. Res.* **2022**, *31*, 693–711. [CrossRef]