Explainable AI-Assisted Theory Generation and Mechanism Identification for Sustainable Tourism: A Methodological Framework

1st Bangxun Zhang
Shell-shaped profile
Wuhu, China
491880732@qq.com

2nd Huan Yu
Shell-shaped profile
Wuhu, China
HuanYu40@outlook.com

3rd Mingding Liang
Shell-shaped profile
Wuhu, China
515385890@qq.com

Abstract*—As research and practice in sustainable tourism continue to evolve, the field increasingly faces a mismatch between rapidly changing practical challenges and the slower pace of theoretical development. Much of the existing literature relies on traditional qualitative approaches — such as Grounded Theory—or descriptive quantitative analysis. While valuable, these methods often struggle to systematically generate new theoretical insights from large-scale, multi-source datasets. Their limitations become particularly evident when attempting to capture the dynamic, nonlinear, and highly interconnected nature of tourism systems, where multiple social, environmental, and behavioral factors interact in complex ways. To address these challenges, this paper proposes a methodological framework that integrates Explainable Artificial Intelligence (XAI) to support theory generation and mechanism identification in sustainable tourism research. The framework leverages widely accessible data sources and cost-effective analytical tools, combining the strong pattern-recognition capacity of machine learning models with causal inference techniques. Through this integration, it establishes a structured pathway that moves from data-driven discovery to theoretically meaningful explanation, enabling researchers to uncover underlying mechanisms rather than merely identifying correlations. To demonstrate the practical application of the framework, the paper uses the formation mechanism of tourists' environmentally responsible behavior as an illustrative case. The analysis integrates multi-source heterogeneous data, including tourist reviews, social media activity, and geospatial information. Within this framework, XAI methods such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) are employed to interpret the internal decision logic of complex predictive models, including deep learning networks. In addition, causal discovery algorithms, such as the PC algorithm and the Fast Causal Inference (FCI) algorithm, are applied to explore potential causal relationships among the identified variables. The results show that XAI not only improves the transparency and interpretability of predictive models, but also plays a crucial role in revealing the key driving factors and their interactive relationships underlying sustainable tourism behaviors. By identifying how different variables influence outcomes and how they interact with one another, the framework helps explain the "why" and "how" behind observed tourism phenomena. This process enables researchers to move beyond simple correlation analysis and toward the development of theoretical constructs and propositions with causal significance, thereby facilitating the emergence of new theoretical insights. Overall, the methodological framework proposed in this study offers a new research paradigm for sustainable tourism studies. By bridging the gap between the predictive strength of big-data analytics and the explanatory depth required for theory development, the framework accelerates the process of theoretical innovation in the field. At the same time, it provides tourism destination managers and policymakers with more precise, evidence-based, and forward-looking decision-support tools, ultimately contributing to the advancement of sustainable tourism development.

Keywords—Explainable AI; Sustainable Tourism; Theory Generation; Mechanism Identification; Grounded Theory

I. INTRODUCTION

Sustainable tourism has emerged as an important pathway for addressing global environmental change and socio-economic challenges, and it has become a key focus in both tourism research and industry practice [1][2]. Despite the growing attention it receives, theoretical development in this field still faces notable difficulties. Tourism systems are inherently complex, dynamic, and highly context-dependent, which makes it challenging for conventional research approaches to uncover their deeper mechanisms [3].

On one hand, tourism scholarship has often been criticized for its limited generation of original theory, with many studies relying on borrowed theoretical frameworks from disciplines such as sociology, economics, and psychology [4]. On the other hand, commonly used theory-building approaches—such as the Grounded Theory Method (GTM)—have important strengths in qualitative analysis but encounter limitations when dealing with the large-scale, multi-modal, and high-dimensional datasets increasingly available in tourism research. These methods are not always well suited to capturing the nonlinear relationships and complex interdependencies that characterize tourism systems [5].

At the same time, Artificial Intelligence (AI) — particularly Machine Learning (ML) — is being adopted rapidly across the tourism sector. From personalized travel recommendations and demand forecasting to operational management and service optimization, AI technologies are transforming how tourism systems operate and are studied [6]. However, many advanced ML models, especially deep learning architectures, function as so-called "black boxes." Their internal decision-making processes are often opaque, making it difficult to interpret how predictions are generated.

This lack of transparency presents two major challenges. First, it reduces the credibility and practical usability of AI models in important decision-making contexts. Second, and

Corresponding Author: Huan Yu, No. 38, Gangwan Road, Economic and Technological Development Zone, Wuhu, China, 241006, HuanYu40@outlook.com

more importantly for academic research, it prevents scholars from extracting meaningful theoretical insights from these models. In other words, researchers may know what a model predicts, but they cannot easily explain why those predictions occur or what underlying causal mechanisms may be driving them. This gap between predictive power and explanatory understanding has become a key limitation in AI-driven tourism research and a significant barrier to theoretical innovation [7].

To address this challenge, the concept of Explainable Artificial Intelligence (XAI) has gained increasing attention. XAI focuses on making machine learning models more transparent by providing techniques that reveal how models reach their decisions and which factors influence their outputs [8]. In recent years, XAI has demonstrated considerable value in high-stakes domains such as finance, healthcare, and autonomous systems, where interpretability and trust are essential.

However, the use of XAI in tourism research, particularly in the context of theory development for sustainable tourism, remains relatively limited [9]. Existing studies primarily employ XAI methods to improve transparency in specific predictive tasks — for example, explaining models used to forecast tourist satisfaction or hotel demand [10][11]. Systematic investigations into how XAI can contribute to the generation of new theoretical constructs and the identification of underlying mechanisms are still scarce.

Against this background, the primary objective of this study is to develop a methodological framework that integrates Explainable AI to support theory generation and mechanism identification in sustainable tourism research. Specifically, the study seeks to address the following central question:

How can XAI be systematically applied to help researchers discover new theoretical constructs, formulate theoretical propositions, and reveal the causal mechanisms underlying tourism phenomena based on complex tourism data?

Importantly, this research does not aim to propose a new substantive theory of sustainable tourism. Instead, it focuses on developing an innovative methodological approach — a "theory of theory building." By positioning XAI not merely as a technical analytical tool but as a conceptual bridge connecting data, computational models, and theoretical development, this study aims to help close the gap between the rapid growth of data analytics capabilities and the comparatively slow pace of theoretical advancement in sustainable tourism research.

The remainder of the paper is organized as follows. First, it reviews the current state of research on theoretical development in sustainable tourism and the application of XAI in related fields. Next, it introduces the proposed methodological framework and its core components. The paper then illustrates the potential application of the framework through a conceptual case example. Finally, it discusses the theoretical contributions, practical implications, limitations, and future research directions of this approach.

II. Literature Review

The theoretical foundation of this study lies at the intersection of three interdisciplinary domains: the theoretical development of sustainable tourism, traditional theory-building methodologies, and Explainable Artificial Intelligence (XAI). Reviewing the literature across these areas helps clarify both the achievements and limitations of existing research, thereby highlighting the necessity and innovative contribution of the framework proposed in this study.

A. The Dilemma of Theoretical Development in Sustainable Tourism

Since the emergence of the concept, the development of a coherent theoretical framework for sustainable tourism has remained a central concern in tourism research. Early studies primarily focused on conceptual clarification, normative principles, and framework construction, which led to the widely recognized "three pillars" model of sustainability — economic, social, and environmental dimensions [12].

However, as sustainable tourism practices have evolved, scholars increasingly recognize that tourism systems operate as Complex Adaptive Systems (CAS). These systems are characterized by nonlinear relationships, dynamic interactions, and cross-scale coupling among multiple actors and environmental factors [3]. Such characteristics make it difficult for traditional theoretical models — often based on linear assumptions and simplified causal relationships — to adequately explain real-world tourism dynamics.

In response to this challenge, researchers have explored a variety of methodological approaches. Among them, the Grounded Theory Method (GTM) has become one of the most widely applied frameworks for theory generation in tourism research [5]. As a bottom-up qualitative methodology, GTM enables researchers to generate theoretical insights directly from empirical data through processes such as systematic coding, constant comparison, and conceptual abstraction. It has been particularly useful for exploring micro-level phenomena such as tourist experiences, community participation, and stakeholder relationships.

Despite these contributions, the limitations of GTM are becoming increasingly apparent. First, the approach relies heavily on the researcher's theoretical sensitivity and subjective interpretation, which makes it difficult to standardize and replicate research procedures. Second, GTM is primarily designed for small-sample qualitative data, such as interview transcripts or field observations. As a result, it struggles to process the massive, multi-modal, and unstructured datasets generated by contemporary tourism activities, including social media content, sensor data, and online platform records. Third, the theories produced through GTM are often context-specific substantive theories, which may have limited generalizability and predictive capacity, making them difficult to apply directly to macro-level policy design or large-scale industry decision-making [5].

As Stumpf et al. have noted, the application of GTM in tourism research has not always fulfilled its full theory-building potential; many studies remain descriptive rather than explanatory, failing to progress toward higher levels of theoretical abstraction [5].

B. The Application of AI in Tourism Research and the "Black Box" Problem

With the rapid advancement of Artificial Intelligence (AI) technologies — especially machine learning and deep learning — tourism research has entered a new era of data-driven analysis. Scholars have increasingly adopted AI-based models to process large-scale datasets and investigate issues such as tourist behavior prediction, tourism demand forecasting, hotel pricing strategies, and sentiment analysis of online reviews [6][13].

These approaches have significantly improved the accuracy and efficiency of empirical analysis, enabling researchers to study tourism phenomena at scales that were previously impossible. For instance, deep learning techniques can analyze millions of online reviews to identify key drivers of tourist satisfaction, while mobile signaling data can reveal fine-grained spatiotemporal patterns of tourist mobility [14].

However, the impressive predictive power of AI models comes with a major challenge—the "black box" problem. Many high-performing algorithms, such as deep neural networks and gradient boosting models, possess complex internal structures and opaque decision-making processes. Although these models may produce highly accurate predictions, researchers often struggle to understand how and why the models reach particular conclusions.

This situation leads to what might be described as "knowing what, but not knowing why." As a result, AI applications in tourism research often remain at the level of technical tools, providing useful predictions without generating deeper theoretical insights. For example, a machine learning model might identify which tourists are more likely to choose eco-tourism products, but it may not explain why certain factors influence those choices or how different variables interact to shape decision-making.

Consequently, the lack of interpretability has limited the potential of AI to contribute to theoretical advancement in tourism research and may even widen the gap between empirical analysis and theory development [7].

C. The Emergence and Application of Explainable AI (XAI)

To address the limitations associated with black-box models, the field of Explainable Artificial Intelligence (XAI) has rapidly developed. XAI does not represent a single technique; rather, it refers to a collection of methods designed to improve the transparency, interpretability, and trustworthiness of AI systems [8].

From an analytical perspective, XAI approaches can generally be divided into global explanations and local explanations.

Global explanations aim to reveal the overall behavior and logic of a model. For instance, SHAP (SHapley Additive exPlanations) values quantify the average contribution of each feature to the model's predictions across the entire dataset [15].

Local explanations, in contrast, focus on explaining individual predictions. LIME (Local Interpretable Model-agnostic Explanations), for example, interprets a single prediction by approximating the model's behavior with a simpler and more interpretable surrogate model in the vicinity of the prediction instance [16].

In recent years, XAI has begun to appear in tourism research. Some studies have used XAI techniques to interpret predictive models related to tourist satisfaction [10], hotel booking cancellations [11], and tourism demand forecasting [17]. These applications improve not only the credibility of machine learning models but also their practical value for industry decision-making.

For example, using XAI, tourism managers can move beyond simply identifying factors associated with high satisfaction and instead understand which specific service elements — such as front-desk interactions or room cleanliness—contributed to negative reviews.

More recently, emerging research has explored combining XAI with causal inference techniques, attempting to uncover causal relationships rather than mere correlations within complex datasets [18][19]. This development opens new possibilities for moving beyond model explanation toward mechanism identification in tourism systems.

D. Research Gap and Positioning of This Study

Taken together, the literature reveals a clear research gap. On one hand, sustainable tourism research urgently requires new theory-building methodologies capable of handling complexity and extracting knowledge from large-scale datasets. On the other hand, although AI technologies provide powerful data-processing capabilities, their black-box nature limits their usefulness for generating theoretical insights. Meanwhile, the emerging application of XAI has largely focused on post-hoc explanations of predictive models, rather than developing a systematic framework for assisting theory generation and mechanism discovery [20].

In essence, traditional grounded theory offers the philosophical logic of inductive theory generation, but lacks the computational tools necessary for analyzing big data. Conversely, modern AI technologies provide powerful data mining capabilities, yet lack a clear pathway for translating patterns into theoretical knowledge [21].

This study seeks to build a bridge between these two domains. We argue that XAI should not be viewed merely as a technical solution for improving model transparency. Instead, it should be recognized as a potential engine for scientific discovery. By combining the bottom-up inductive logic of grounded theory with the pattern-recognition power of machine learning, XAI can enable what may be termed an "Augmented Grounded Theory" approach.

Accordingly, this paper aims to address the following central methodological question:

- How can a research framework centered on XAI guide scholars in transforming complex tourism data into theoretical knowledge — progressing systematically from data to patterns, from patterns to explanations, and ultimately from explanations to theory?

Answering this question represents the core methodological challenge that this study seeks to address and a critical step toward advancing theory development in sustainable tourism research.

III. Methodology: An Explainable AI-Assisted Theory Generation Framework

To systematically apply Explainable Artificial Intelligence (XAI) in promoting theoretical discovery within sustainable tourism research, this study proposes a four-stage framework termed "XAI-Assisted Theory Generation" (XAI-TG). The core idea of this framework is to integrate the bottom-up inductive logic of traditional Grounded Theory (GTM) with the data-driven analytical and explanatory capabilities of XAI, thereby forming what can be described as an Augmented Grounded Theory approach. Through this integration, the framework aims to enable a systematic and reproducible transition from data to theoretical insight. Rather than focusing solely on prediction accuracy, the framework emphasizes explaining why phenomena occur and identifying the mechanisms through which they operate, thus providing stronger empirical foundations for theory development.

A. Research Strategy: An Augmented Grounded Theory Path

The XAI-TG framework follows an iterative research strategy. Its overall structure parallels the classical open coding – axial coding – selective coding process of grounded theory, while each stage is enhanced through the use of AI and XAI techniques (Figure 1).

1) Phenomenon Identification and Data Fusion (Corresponding to Open Coding)

The process begins with the identification of a specific and meaningful sustainable tourism phenomenon, such as tourists' environmentally responsible behavior. Once the research phenomenon is defined, the next step is to collect and integrate multi-source datasets that reflect different dimensions of the phenomenon. These datasets may include online reviews, social media content, geospatial information, and socio-economic statistics.

The purpose of this stage is to construct a rich, multidimensional data environment capable of capturing the complexity of the tourism phenomenon under investigation. Similar to the open coding stage in grounded theory, the objective here is to identify relevant variables and signals within the data that may later form the basis for theoretical constructs.

2) Predictive Modeling and Pattern Discovery (Corresponding to Axial Coding)

In the second stage, researchers develop predictive models to analyze the relationship between explanatory variables and a key outcome variable. To ensure interpretability and replicability, relatively transparent models — such as decision trees or logistic regression models—are preferred.

It is important to emphasize that the objective at this stage is not to achieve maximum predictive accuracy. Instead, the model serves as a "computational lens" that helps reveal complex nonlinear relationships and higher-order interaction patterns embedded in the data—patterns that may be difficult for researchers to detect through conventional statistical analysis.

A predictive model that successfully explains or predicts a tourism phenomenon implicitly contains encoded information about the underlying structure and dynamics of that phenomenon. These encoded relationships become the starting point for deeper explanatory analysis.

3) XAI Explanation and Construct Refinement (Corresponding to Axial Coding)

This stage represents the core component of the XAI-TG framework. Researchers employ a variety of XAI techniques, such as SHAP, LIME, or attention-based explanation methods, to interpret the predictive model developed in the previous stage.

Through both global explanations and local explanations, researchers can identify the most influential variables affecting model predictions, which may correspond to potential theoretical constructs. XAI methods also allow researchers to quantify:

- The relative importance of different variables
- The direction and magnitude of their effects
- The interaction relationships among variables

For example, SHAP dependence plots may reveal nonlinear relationships between variables — for instance, an inverted U-shaped relationship between income level and tourists' environmental responsibility. Meanwhile, SHAP interaction values may highlight synergistic relationships between variables, such as the combined effect of high education levels and frequent visits to natural parks on environmentally responsible behavior.

The stable patterns revealed at this stage serve as proto-theoretical propositions, providing insight into the relationships that may underpin broader theoretical explanations.

4) Causal Inference and Mechanism Validation (Corresponding to Selective Coding)

While XAI methods reveal patterns and associations, correlation alone is insufficient for robust theory development. Therefore, the final stage of the framework aims to identify causal relationships underlying the patterns discovered in earlier stages.

Researchers may employ statistical causal inference techniques — such as regression analysis, propensity score matching, or causal discovery algorithms — to explore the causal relationships among variables. In addition, experimental or quasi-experimental designs can be used to validate specific hypotheses derived from the XAI analysis.

For example, if XAI results suggest that destination environmental information disclosure strongly influences tourists' environmentally responsible behavior, researchers could design an online experimental study. Different groups of participants might be presented with varying levels of environmental information about a destination, allowing researchers to observe whether this information causally influences booking decisions or behavioral intentions.

Through this process, theoretical propositions can evolve from simple associations ("A is related to B") to causal explanations ("A leads to B through specific mechanisms").

5) Framework Outcome

By integrating these four interconnected stages, the XAI-TG framework provides a structured pathway for researchers

to move from raw data to theoretical explanation. The framework ensures that emerging theories remain empirically grounded in data, while also progressing beyond descriptive patterns toward mechanistic and causal understanding.

In this way, XAI-TG offers a systematic methodology for theory generation in data-rich environments, helping bridge the gap between big data analytics and theoretical innovation in sustainable tourism research.

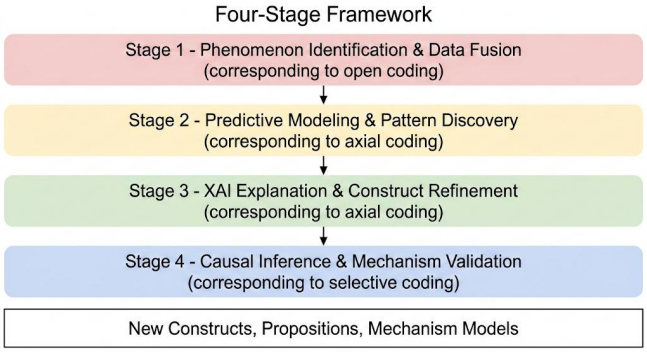

Fig. 1. XAI-Assisted Theory Generation (XAI-TG) Framework]

B. *Core Technologies of the Framework: XAI and Causal Inference*

The innovativeness of this framework lies in its integrated application of XAI and causal inference techniques. The key technologies are briefly explained below (Table I).

TABLE I. CORE TECHNOLOGIES IN THE XAI-TG FRAMEWORK AND THEIR ROLES IN THEORY GENERATION

Technology Category	Specific Technology (Examples)	Role in Theory Generation
Global Explanation Techniques	SHAP (SHapley Additive exPlanations)	Identify Core Constructs: Identify the most influential key drivers of the phenomenon by calculating the average marginal contribution of all features. Discover Non-linear Relationships: Reveal non-linear patterns (e.g., U-shaped, threshold effects) between individual variables and the outcome.
Local Explanation Techniques	LIME (Local Interpretable Model-agnostic Explanations)	Understand Heterogeneity: Explain the model's prediction logic for specific individuals or small groups to discover mechanism differences in different contexts. Diagnose Anomalies and Novel Patterns: Attribute the model's "unexpected" predictions, potentially revealing unanticipated influence paths or special groups.
Interaction Effect Analysis	SHAP Interaction Values, Friedman's H-statistic	Establish Theoretical Propositions: Identify synergistic or antagonistic effects between different factors, forming complex propositions such as "if A and B exist simultaneously, then C will be enhanced/weakened."
Causal Discovery Algorithms	PC (Peter-Clark) Algorithm, FCI (Fast Causal Inference)	Construct Causal Structure: Learn the causal relationship graph among variables from observational data, providing a structural hypothesis for mechanism identification.
Causal Effect Estimation	Difference-in-Differences (DID), Regression Discontinuity Design (RDD), Propensity Score Matching (PSM)	Validate Causal Mechanisms: Quantify the net causal effect of a specific intervention (factor) on the outcome in quasi-experimental or observational studies, confirming theoretical hypotheses.

C. *Data Processing and Model Building*

In the application of the XAI-Assisted Theory Generation (XAI-TG) framework, data serves as the fundamental basis for theoretical discovery. Researchers need to build a multi-source, heterogeneous database tailored to the specific research question being investigated. Because tourism phenomena are often reflected in diverse forms of data — such as text, spatial records, and behavioral traces — data preprocessing becomes a critical step in ensuring the quality and usability of the dataset.

1) Text Data Processing

Tourism-related textual data, including tourist reviews, travel blogs, and social media posts, often contain rich information about tourists' perceptions, experiences, and behavioral intentions. To transform these unstructured texts into usable analytical variables, researchers can apply Natural Language Processing (NLP) techniques. Typical procedures include:

- Word segmentation and tokenization to identify meaningful textual units
- Sentiment analysis to quantify emotional polarity and intensity
- Topic modeling, such as Latent Dirichlet Allocation (LDA), to identify key thematic patterns within large text corpora

Through these steps, qualitative textual information can be converted into quantifiable features that can be incorporated into predictive models.

2) Spatial Data Processing

Tourism research increasingly utilizes spatial and mobility data, such as tourists' GPS trajectories, geotagged social media posts, and Point of Interest (POI) distributions. These spatial datasets can reveal important behavioral patterns related to tourist movement and destination interaction.

Common spatial feature extraction methods include:

- Activity radius, which measures the spatial extent of tourist movement
- Spatial entropy, reflecting the diversity and dispersion of visited locations
- Stay hotspots, identifying areas where tourists spend extended periods of time

These indicators help capture the spatial dimension of tourist behavior, which is often closely related to environmental attitudes and sustainable travel practices.

3) Data Fusion

Because tourism datasets often originate from multiple sources, an essential step is data fusion, which involves aligning and integrating different datasets into a unified analytical structure. This process requires matching information across datasets according to a common unit of analysis, such as:

- an individual tourist
- a specific destination
- a defined time period

By harmonizing heterogeneous data sources, researchers can create a comprehensive and multidimensional dataset capable of supporting more sophisticated analytical models.

4) Model Selection and Evaluation

When selecting predictive models within the XAI-TG framework, priority should be given to models that balance interpretability, performance, and cost-effectiveness. Such models are more practical for real-world research and decision-making contexts. Importantly, a model with strong predictive performance is more likely to capture meaningful patterns and relationships present in real-world tourism systems, thereby providing a reliable basis for subsequent XAI analysis.

Model evaluation should therefore extend beyond traditional performance indicators such as:

- Accuracy
- Precision
- Recall
- F1-score

In addition, the framework recommends incorporating interpretability-oriented evaluation metrics, including:

- Fidelity, which measures how accurately the explanation method reflects the behavior of the original model
- Stability, which assesses whether explanations remain consistent when small changes are introduced to the input data

By combining predictive performance evaluation with interpretability assessment, researchers can ensure that the selected models are both analytically robust and theoretically informative.

Overall, careful data preparation, feature extraction, and model evaluation are essential steps in the XAI-TG framework. They ensure that the insights derived from XAI explanations are reliable, meaningful, and capable of supporting the generation of new theoretical constructs in sustainable tourism research.

IV. Results: A Conceptual Validation of the Framework's Application

To illustrate the practical application process and potential analytical outputs of the XAI-TG framework, this section presents a conceptual case titled "Exploring the Key Drivers and Formation Mechanisms of Tourists' Environmentally Responsible Behavior (ERB)." The purpose of this example is not to report a complete empirical study but to demonstrate how the framework operates in practice and what types of insights it can generate. The results presented here should therefore be understood as expected outcomes derived from the methodological logic of the framework.

A. Stage One: Phenomenon Identification and Data Fusion

The research phenomenon is defined as tourists' environmentally responsible behavior (ERB) exhibited at the destination. The outcome variable (Y) is operationalized as a composite ERB score derived from the text and image content shared by tourists on social media platforms.

Using Natural Language Processing (NLP) and computer vision techniques, this score captures observable indicators of environmentally responsible actions, such as:

- references to waste sorting or recycling
- mentions of participation in environmental activities
- evidence of environmentally harmful behavior in uploaded photos
- expressions of environmental awareness or conservation attitudes

To identify the driving forces behind ERB, the study integrates multiple sources of heterogeneous data, including:

1) Tourist Personal Characteristics

Derived from user registration information or surveys:

- age
- gender
- education level
- income level

2) Tourism Behavior Data

Collected from Online Travel Agency (OTA) booking records and GPS trajectories:

- travel frequency
- preferred destination types (urban vs. natural)
- length of stay
- tourism expenditure level

3) Social Network Data

Extracted from tourist-generated social media content (e.g., Weibo, Xiaohongshu):

- sentiment orientation toward environmental issues
- topics of interest related to travel
- social network centrality and peer influence

4) Destination Characteristic Data

Obtained from Geographic Information Systems (GIS) and official statistics:

- environmental quality indicators (e.g., AQI index)

- proportion of green-certified hotels
- availability of environmental facilities
- intensity of environmental promotional campaigns

Through data fusion, these sources are integrated into a large panel dataset containing tens of thousands of tourist observations, with each observation including hundreds of potential explanatory variables (X).

B. Stage Two: Predictive Modeling and Pattern Discovery

In the second stage, a Gradient Boosting Decision Tree model (XGBoost) is used to predict tourists' ERB scores.

After cross-validation and hyperparameter tuning, the model achieves strong predictive performance on the test dataset, with an R^2 value of approximately 0.75. This indicates that the model successfully captures the complex relationships and behavioral patterns associated with tourist ERB.

Although the model functions as a high-performance "black box," its predictive accuracy suggests that it has effectively encoded important information about the underlying drivers of environmentally responsible behavior. This predictive model therefore serves as the analytical foundation for the subsequent XAI explanation stage.

C. Stage Three: XAI Explanation and Construct Refinement

To interpret the XGBoost model and identify theoretical insights, SHAP (SHapley Additive exPlanations) is applied.

1) Identification of Core Drivers (Global Explanation)

By calculating the mean absolute SHAP value for each feature, the analysis identifies the top ten factors that most strongly influence tourist ERB (Figure 2).

These features can be grouped into several potential theoretical constructs:

a) Environmental Awareness

Variables such as:

- frequency of environmental keywords in pre-trip searches
- attention to environmental topics on social media

highlight the central role of individual environmental consciousness.

b) Destination Context

High-ranking features such as:

- intensity of destination environmental promotion
- proportion of green-certified hotels

indicate that institutional and environmental contexts strongly influence tourist behavior.

c) Social Norms

Variables including:

- average ERB level within the tourist's peer group
- proportion of environmentally responsible behaviors among social network friends

demonstrate the influence of social interaction and normative pressure.

d) Perceived Personal Cost

Factors negatively affecting ERB include:

- additional time required for eco-friendly transportation
- price premium for green hotels

These results highlight the role of economic and convenience-related constraints.

2) Discovery of Nonlinear Relationships and Interaction Effects

SHAP dependence plots further reveal complex nonlinear and interaction effects between variables and ERB (Figure 3).

a) Threshold Effects

The effect of destination environmental promotion intensity on ERB remains relatively weak at low levels but increases dramatically once a critical threshold is reached.

b) Inverted U-Shaped Relationship

The relationship between travel frequency and ERB shows an inverted U-shaped pattern:

- tourists who travel very rarely show low ERB engagement
- tourists with moderate travel frequency demonstrate the highest ERB levels
- extremely frequent travelers exhibit declining ERB

This pattern may reflect diminishing marginal environmental sensitivity or experience fatigue.

c) Interaction Effects

SHAP interaction analysis reveals a strong synergistic relationship between education level and environmental promotion intensity.

In other words, environmental campaigns are significantly more effective among highly educated tourists.

This finding supports the theoretical proposition that:

"The effectiveness of environmental communication is moderated by tourists' cognitive capacity."

D. Stage Four: Causal Inference and Mechanism Validation

While XAI reveals strong associations between variables, the final stage aims to identify underlying causal mechanisms.

1) Construction of a Causal Hypothesis Graph

Combining XAI findings with existing theoretical insights, the PC causal discovery algorithm is used to infer a preliminary causal structure (Figure 4).

The resulting model suggests that:

- Education level and social network information exposure influence environmental awareness
- Environmental awareness and destination context jointly shape perceived personal cost
- these factors ultimately determine environmentally responsible behavior

2) Quasi-Experimental Mechanism Validation

To test the causal impact of destination environmental promotion intensity, a quasi-experimental analysis using Propensity Score Matching (PSM) is conducted.

Two groups of tourists are identified:

- Treatment group: exposed to high-intensity environmental promotional information
- Control group: not exposed to such information

Each treated individual is matched with a statistically similar control individual based on observable characteristics.

The results show that:

- the average ERB score of the treatment group is significantly higher than that of the control group
- the estimated effect size is $d = 0.45$ ($p < 0.001$)

Further mediation analysis indicates that environmental awareness explains approximately 30% of the total effect, confirming the mechanism:

Environmental Promotion → Increased Environmental Awareness → Higher ERB

E. Framework Outcome

Through these four analytical stages, the XAI-TG framework enables researchers to:

- identify key drivers of tourist ERB
- reveal nonlinear and interaction effects
- uncover causal mechanisms underlying behavior

The process ultimately generates a set of testable theoretical propositions and mechanism models. These outputs provide a strong empirical foundation for developing a comprehensive theory of environmentally responsible tourist behavior, while also demonstrating the practical value of XAI in supporting data-driven theory generation in sustainable tourism research.

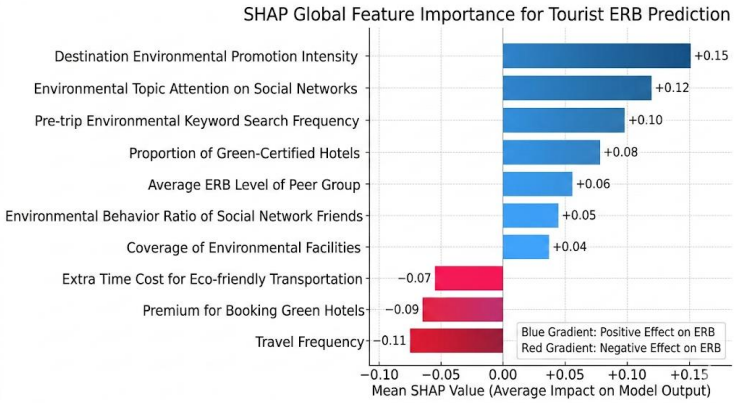

Fig. 2. SHAP Global Feature Importance for Tourist ERB Prediction

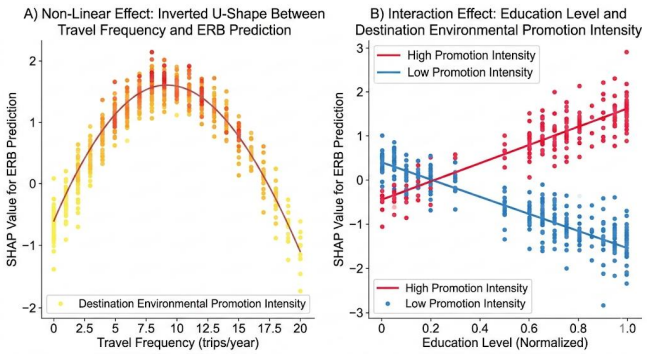

Fig. 3. SHAP Dependence Plots Revealing Non-linear and Interaction Effects

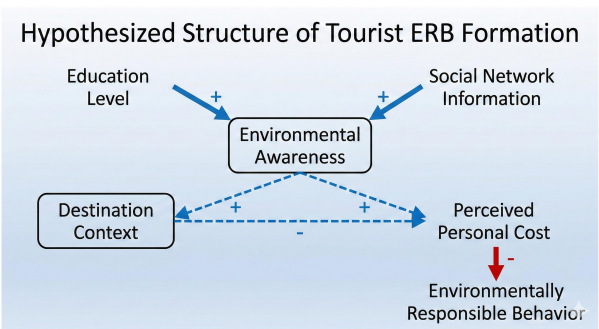

Fig. 4. Hypothesized Causal Structure of Tourist ERB Formation]

V. Discussion

The XAI-Assisted Theory Generation (XAI-TG) framework is intended to offer a new paradigm for sustainable tourism research by explicitly linking big data analytics with theory building. Below is an in-depth discussion of its core value, how it differs from existing approaches, its theoretical and practical contributions, and key limitations and future directions.

A. Core Value: From "Prediction" to "Explanation" to "Understanding"

Much data-driven tourism research — especially work built on high-performing black-box models — typically culminates in prediction. These studies can forecast demand, segment tourists, or predict behaviors with impressive accuracy, yet often struggle to explain why the phenomenon occurs or how its mechanisms operate.

The distinctive value of XAI-TG is that it systematically constructs a cognitive upgrading pathway:

1) Prediction (pattern localization):

High-performance models are used as a kind of prospecting instrument to detect complex, nonlinear patterns and higher-order interactions within massive datasets — signals that are difficult to find through human reading or classical linear modeling.

2) Explanation (human-readable structure):

XAI methods function like a scalpel, opening the model to reveal which variables matter most, how they matter (direction and magnitude), and when they matter (nonlinearities, thresholds), as well as how factors interact. This transforms a purely predictive artifact into an interpretable map of relationships.

3) Understanding (mechanism and causality):

Most crucially, XAI-TG introduces causal inference to move beyond "explaining correlations" toward testing causal hypotheses and mechanism pathways. This step is what allows the research to approach deeper understanding of system operation, rather than remaining at the level of descriptive explanation.

This layered logic directly targets the long-standing bottleneck in sustainable tourism: theory development lags behind the complexity of practice. XAI-TG reframes AI from a "prediction tool" into a partner for theoretical discovery, aligning with GTM's philosophy of letting theory emerge from data—while equipping that philosophy with a more efficient, scalable, and reproducible analytical engine. In this sense, XAI-TG can be reasonably framed as a computationally augmented grounded theory pathway.

B. Dialogue with Existing Research: Beyond GTM and Beyond "Black-Box" AI

1) Beyond traditional Grounded Theory (GTM)

GTM's strength lies in deep contextualization and concept generation through iterative coding. Yet the method faces clear constraints under contemporary tourism big data conditions:

- Scale and modality constraints: GTM is typically designed for small-to-moderate qualitative datasets (interviews, observations). It is not naturally suited to integrating and mining multi-modal data such as reviews, images, mobility traces, and platform logs at large scale.
- Labor intensity and efficiency: Manual coding is time-consuming and becomes infeasible when data volumes grow by orders of magnitude.
- Subjectivity and reproducibility: Heavy reliance on researcher interpretation makes standardization and replication difficult, even with careful audit trails.

XAI-TG addresses these limits by enabling automated pattern discovery over large heterogeneous datasets, detecting weak signals, nonlinearities, and high-order interactions that human coders might miss. Moreover, XAI outputs (e.g., SHAP-based rankings and dependence patterns) can provide a more explicit and transparent basis for construct prioritization, partially mitigating researcher bias and enhancing procedural reproducibility.

2) Beyond mainstream "black-box" AI in tourism

Most AI applications in tourism are evaluated primarily through task performance (accuracy, RMSE, AUC). Their major academic contribution often becomes: "we improved prediction." But this leaves a persistent "so what?" problem when the goal is theory development.

XAI-TG directly confronts this explanatory gap by asserting that a model's value lies not only in its outputs but in the knowledge it encodes about how a phenomenon operates. The framework formalizes a method for extracting that knowledge — turning AI from a technical tool into a methodological innovation for scientific discovery. This is the key shift: from "using AI to predict tourism outcomes" to "using AI to build theoretical explanations about tourism systems."

C. Theoretical and Practical Contributions

1) Theoretical contributions

The main theoretical contribution of XAI-TG is methodological: it provides a systematic, data-driven theory generation pathway for sustainable tourism and, by extension, other social science domains confronting complex systems and large-scale data.

Key advances include:

- Integration of three frontiers: machine learning (pattern discovery), XAI (interpretable explanations), and causal inference (mechanism testing) are integrated into a single coherent workflow.
- Paradigm evolution: rather than relying solely on deductive hypothesis testing or purely qualitative induction, XAI-TG supports a discovery cycle where induction and deduction iterate: patterns suggest constructs; constructs form propositions; propositions are tested and refined.
- Operationalization of "Augmented Grounded Theory": it offers a concrete roadmap for bringing GTM's "emergence from data" into the big-data era without abandoning rigor.

2) Practical contributions

For destination managers and policymakers, the framework's value lies in translating analytics into operationally actionable mechanism insights. Traditional predictive analytics might say: "Factors A, B, and C correlate with satisfaction/ERB." XAI-TG can go further by supporting:

- Importance ranking: which drivers should be prioritized in resource allocation?
- Mechanisms of action: is the effect linear or threshold-based? where are the tipping points?
- Heterogeneous effects: which groups respond differently (market segments, socioeconomic strata, travel styles)?
- Synergies and trade-offs: do combined interventions yield "1+1>2" effects, or do they conflict?

These insights enable precision governance and more targeted sustainable tourism strategies, especially in contexts where interventions must be efficient, politically feasible, and sensitive to stakeholder heterogeneity.

D. Limitations and Future Directions

Despite its promise, XAI-TG also has clear limitations that should be acknowledged.

1) Limitations

a) Strong dependence on data quality and coverage

If data are biased, incomplete, or omit key variables, any discovered "drivers" may reflect sampling artifacts rather than real mechanisms. The framework does not remove the fundamental "garbage in, garbage out" constraint—it can even amplify it if researchers over-trust outputs.

b) High technical threshold and interdisciplinary demands

Effective implementation requires combined competence in tourism theory, machine learning, XAI tooling, and causal inference. This is a significant barrier for many research teams trained primarily in traditional social science methods.

c) Explanation ≠ causation

Most XAI outputs are still fundamentally association-based attribution. Even causal discovery methods applied to observational data should be treated as generating causal hypotheses, not final truths. Robust causal validation still requires careful design — experiments, quasi-experiments, natural experiments, or strong identification strategies.

d) Risk of "mechanism over-interpretation"

Rich explanation visuals and rankings can create an illusion of certainty. Without triangulation, researchers may prematurely convert model artifacts into theoretical claims.

2) Future directions

a) Empirical applications across diverse sustainable tourism topics

Applying XAI-TG to different questions — tourism poverty alleviation mechanisms, destination resilience formation, over-tourism governance, cultural heritage commercialization – protection tensions — would stress-test the framework and help refine best practices.

b) Methodological expansion for deeper mechanism insight

Integrating more advanced XAI and causal methods — such as counterfactual explanations, causal mediation/moderation analysis, invariant causal prediction, or causal representation learning — could strengthen the "understanding" step and reduce the risk of mistaking correlation for mechanism.

c) Tool development to lower barriers

Developing low-code/no-code pipelines (or modular open-source toolkits) that encapsulate data fusion, model training, XAI reporting, and causal testing workflows would greatly increase accessibility and facilitate replication across research teams.

VI. Conclusion

Against the growing demand for deeper theoretical guidance and effective practical pathways in sustainable tourism research, this study addresses a fundamental challenge: the pace of theoretical development has not kept up with the rapid expansion of available data. In response, it proposes an innovative methodological framework that integrates Explainable Artificial Intelligence (XAI) — the XAI-Assisted Theory Generation (XAI-TG) framework. The purpose of this framework is to systematically support researchers in generating new theories and identifying operational mechanisms from complex, multi-source tourism big data, thereby bridging the long-standing gap between predictive capability and explanatory understanding.

The central conclusion of this study is that by combining three complementary analytical capabilities — the pattern discovery power of machine learning, the interpretability provided by XAI, and the mechanism validation offered by causal inference—it is possible to establish a more efficient, transparent, and theoretically meaningful pathway from "data" to "theory." The proposed XAI-TG framework operationalizes this process through four iterative stages:

- Phenomenon identification and data fusion
- Predictive modeling and pattern discovery
- XAI-based explanation and construct refinement
- Causal inference and mechanism validation

Through these stages, the framework integrates the inductive logic of traditional grounded theory with the computational capabilities of modern data science, forming what can be described as an "Augmented Grounded Theory" approach suited to the big-data era.

The primary contribution of this research lies in its methodological innovation. The study offers a forward-looking framework not only for sustainable tourism research but also for the broader social sciences. It demonstrates how artificial intelligence can move beyond its conventional role as a predictive tool and instead function as a collaborative partner in theoretical discovery. By illustrating how XAI can help identify key theoretical constructs, uncover complex relationships among variables, and support causal mechanism exploration, the study provides researchers with a clear and operational roadmap for theory generation in data-rich environments. At the same time, the framework offers tourism managers and policymakers evidence-based insights that can support the development of more targeted, efficient, and sustainable tourism strategies.

Nevertheless, the proposed framework should not be viewed as a universal solution. Its effectiveness depends strongly on the quality, completeness, and representativeness of the available data, as well as on the interdisciplinary expertise of the research team. Moreover, although the framework incorporates causal inference techniques, conclusions about causality ultimately require validation through rigorous experimental or quasi-experimental designs.

Future research should therefore focus on applying and refining the XAI-TG framework across diverse empirical contexts, such as tourism resilience, community-based tourism development, or environmental governance in tourism destinations. At the same time, further methodological advancements — such as integrating more advanced XAI techniques or developing user-friendly analytical tools that lower technical barriers — will be essential. By doing so, the research community can help promote data-driven, AI-enabled theoretical innovation and scientific discovery, advancing both the academic understanding and practical implementation of sustainable tourism.

References

[1] Gössling, S., & Mei, X. Y. (2025). AI and sustainable tourism: an assessment of risks and opportunities for the SDGs. Current Issues in Tourism, 1-18. DOI: 10.1080/13683500.2025.2477142

[2] Han, H., Chua, B. L., & Fakfare, P. (2024). Green marketing: Consumption and development of sustainable tourism and hospitality. Journal of Travel & Tourism Marketing, 41(3), 275-289.

[3] Farrell, B. H., & Twining-Ward, L. (2004). Reconceptualizing tourism. Annals of Tourism Research, 31(2), 274-295.

[4] Franklin, A., & Crang, M. (2001). The trouble with tourism and travel theory?. Tourist studies, 1(1), 5-22.

[5] Stumpf, T. S., Sandstrom, J., & Swanger, N. (2016). Bridging the gap: grounded theory method, theory development, and sustainable tourism research. Journal of Sustainable Tourism, 24(12), 1691-1708.

[6] Wang, S., et al. (2025). Artificial Intelligence in Tourism: A Systematic Literature Review. Sustainability, 17(20), 9080.

[7] Contessi, D., et al. (2024). Decoding the future: Proposing an interpretable machine learning model for occupancy forecasting. International Journal of Hospitality Management, 118, 103684.

[8] Carloni, G., Berti, A., & Colantonio, S. (2025). The role of causality in explainable artificial intelligence. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, e1505.

[9] Kose, U., & Kose, C. (2024). Stakeholder perspectives on integrating explainable artificial intelligence in medical tourism. Tourism and Hospitality Management, 30(1), 123-145.

[10] satisfaction for local Korean festivals using explainable AI. Sustainability, 13(19), 10901.

[11] Huang, L., Zheng, W., Zou, S., & Chen, M. H. (2025). Interpretable machine learning for hotel demand prediction: A case study framework from Xiamen, China. Tourism Economics, 13548166251360295.

[12] WCED (World Commission on Environment and Development). (1987). Our Common Future. Oxford University Press.

[13] Li, Y., et al. (2025). Interpretable machine learning for predicting and explaining building energy consumption. Energy and Buildings, 301, 113721.

[14] Si, R., et al. (2025). Interpretable Machine Learning Insights into the Factors Influencing Intra-Urban Travel Patterns. ISPRS International Journal of Geo-Information, 14(1), 39.

[15] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in neural information processing systems (pp. 4765-4774).

[16] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).

[17] Huang, L., Zheng, W., & Deng, Z. (2024). Tourism demand forecasting: An interpretable deep learning model. Tourism Analysis, 29(4), 535-549.

[18] Karimi, A. H., et al. (2025). Position: Explainable AI is Causal Discovery in Disguise. ICLR 2025 Workshop on Causal Learning.

[19] van Veen, K., et al. (2025). XAI In Fraud Detection: A Causal Perspective. In Machine Learning and Principles and Practice of Knowledge Discovery in Databases (pp. 225-240). Springer.

[20] Spirtes, P., Glymour, C. N., & Scheines, R. (2000). Causation, prediction, and search. MIT press.

[21] Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41-55. University Science, 1989.

ACKNOWLEDGEMENTS

We would like to express our sincere gratitude to all participants who contributed valuable feedback during the development and refinement of the methodology. We also acknowledge the support from those who helped in the collection and organization of data, as well as the valuable suggestions provided during the conceptualization and interpretation of the research. Additionally, we thank the institutions and organizations that provided resources to facilitate this research. Their contributions have significantly enhanced the quality and depth of this study.

FUNDING

None.

AVAILABILITY OF DATA

Not applicable.

AUTHOR CONTRIBUTIONS

Bangxun Zhang: Conceptualization, Methodology, Formal analysis, Visualization, Writing—original draft.

Huan Yu: Data curation, Validation, Investigation, Writing—review & editing.

Mingding Liang: Supervision, Resources, Project administration, Writing—review & editing.

COMPETING INTERESTS

The authors declare no competing interests.