Routledge
Taylor & Francis Group

Journal of Asian Architecture and Building Engineering

ISSN: 1346-7581 (Print) 1347-2852 (Online) Journal homepage: www.tandfonline.com/journals/tabe20

Multi-model expert – AI triangulated validation of highway landscape classification with zero-shot multimodal large language models

Hangyu Gao , Richard Smardon , Shamsul Abu Bakar , Suhardi Maulan , Jiani Yang , Riyadh Mundher & Yijiang Zhou

To cite this article: Hangyu Gao , Richard Smardon , Shamsul Abu Bakar , Suhardi Maulan , Jiani Yang , Riyadh Mundher & Yijiang Zhou (21 Jul 2026): Multi-model expert – AI triangulated validation of highway landscape classification with zero-shot multimodal large language models, Journal of Asian Architecture and Building Engineering, DOI: 10.1080/13467581.2026.2706274

To link to this article: https://doi.org/10.1080/13467581.2026.2706274

JOURNAL OF ASIAN ARCHITECTURE AND BUILDING ENGINEERING
https://doi.org/10.1080/13467581.2026.2706274

URBAN PLANNING AND DESIGN

OPEN ACCESS Check for updates

Multi-model expert – AI triangulated validation of highway landscape classification with zero-shot multimodal large language models

Hangyu Gao[a], Richard Smardon[b], Shamsul Abu Bakar[a], Suhardi Maulan[a], Jiani Yang[c], Riyadh Mundher[a] and Yijiang Zhou[d]

[a]Department of Landscape Architecture, Faculty of Design and Architecture, Universiti Putra Malaysia, Serdang, Malaysia; [b]Department of Environmental Studies, SUNY College of Environmental Science and Forestry, Syracuse, NY, USA; [c]Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, CA, USA; [d]Department of Architecture, City College, Kunming University of Science and Technology, Kunming, China

ABSTRACT

Highway landscapes influence driver safety, cultural identity, and economic development, yet existing classification approaches rarely quantify label reliability or validate AI automation against expert judgment. This study develops a multi-model expert–AI triangulated validation framework for highway landscape classification using multimodal large language models (MLLMs). From 1,828 images sampled at 250 m intervals along 418 km of Malaysia's North–South Expressway, zero-shot classifications from ChatGPT-5.4 Thinking, Claude Sonnet 4.6, and Gemini 3 Thinking were compared with expert annotations across 16 landscape characters. A seven-category agreement scheme yielded 1,352 high-confidence images (74.0% retention). Claude achieved the highest accuracy (70.3%), followed by ChatGPT (68.8%) and Gemini (67.0%); only the Claude–Gemini gap was significant ($p < 0.05$). All models exceeded 90% retention for morphologically distinct landscapes but disagreed on transitional characters, revealing fragile taxonomic boundaries rather than model error alone. Complementary strengths—Claude in infrastructure-heavy environments, ChatGPT in built-environment characters, Gemini in geomorphologically distinctive features—support triangulation over single-model reliance. The framework reframes validation as mapping a "reliability surface," treating MLLMs as additional raters rather than replacements, and the validated dataset supports corridor-scale mapping and AI–human collaboration in environmental assessment.

ARTICLE HISTORY
Received 26 April 2026
Accepted 15 July 2026

KEYWORDS
Highway landscape classification; landscape character; multimodal large language models; AI-expert validation framework; transportation planning

1. Introduction

Highway corridors catalyse economic development in developing and remote regions (Huijuan et al. 2025), generating increased traffic volumes. At the same time, highways can also be understood as "corridors of human values" that provide travel experience opportunities rather than merely efficient routes between origin and destination (Brown 2003). In recreational contexts, driving itself can become a form of pleasure-oriented travel, with scenic road segments deliberately sought out as part of the experience (Hallo and Manning 2009). Consequently, roadside landscapes assume greater significance in shaping viewers' experiences, perceptions, and well-being (Gao et al. 2025). Highway landscapes encompass a diverse range of natural and cultural settings, from agricultural plains to urban peripheries. Regional characteristics – such as palm plantations in tropical zones or volcanic terrain in geologically active areas – contribute to visual heterogeneity that shapes viewers' perceptual experiences (Gao et al. 2024). These landscapes also express local cultural heritage, reflecting the dynamic tension between preservation and adaptive evolution that underpins authentic regional identity (Martín et al. 2018).

Beyond aesthetic value, highway landscapes have a direct influence on driver behaviour and safety (B. Jiang et al. 2021). Research indicates that monotonous landscape sequences – such as extended agricultural plains or repetitive plantation corridors – significantly increase the risk of "highway hypnosis" and fatigue-related accidents by reducing cognitive stimulation (Larue, Rakotonirainy, and Pettitt 2011; Thiffault and Bergeron 2003; Zainon et al. 2018). Environmental conditions along roadsides contribute to approximately 17% of traffic accidents (Guo et al. 2019). Well-designed landscapes guide attention, mitigate fatigue, and improve safety outcomes. Given these multifaceted dimensions – natural, cultural, and safety-related – systematic classification approaches are essential for effective corridor management and design interventions.

CONTACT Richard Smardon rsmardon@esf.edu Department of Environmental Studies, SUNY College of Environmental Science and Forestry, 1 Forestry Drive, Syracuse, NY 13210, USA
Supplemental data for this article can be accessed online at https://doi.org/10.1080/13467581.2026.2706274.

1.1. Theoretical and methodological foundations of Highway landscape classification

Landscape Character Assessment (LCA) provides a foundational framework by systematically documenting characteristics, spatial distributions, and temporal dynamics (Atik et al. 2017; Terkenli, Gkoltsiou, and Kavroudakis 2021). At its core, landscape character denotes the distinct spatial configurations that differentiate one area from another (Jellema et al. 2009). Each character area embodies a unique assemblage of variables identified through systematic characterisation processes (James and Gittins 2007). The LCA methodology proceeds through a structured sequence: parameter definition, data analysis, field validation, classification refinement, and planning output generation (Tudor 2014). Different criteria yield divergent taxonomic structures, confirming that absolute spatial units do not objectively exist (Bastian 2008). Beyond practical applications, classification serves as a theoretical instrument for standardised communication and knowledge synthesis in landscape science (Brabyn 2009).

In conventional practice, such classifications are predominantly expert-led and follow the four-step Landscape Character Assessment (LCA) process set out by Tudor (2014). Assessors first define the purpose, scope, and scale of the study (Step 1), then conduct a desk study (Step 2) that overlays two broad categories of spatial data: natural factors (geology, landform, hydrology, soils, climate, and land cover) and cultural/social factors (land use, settlement pattern, field enclosure, land ownership, and historic time depth), supplemented by cultural associations from art, literature, and local memory. Draft units of common character are delineated from this overlay analysis. A subsequent field survey (Step 3) verifies and refines these desk-based units by recording aesthetic attributes – scale, enclosure, diversity, texture, form, colour, pattern – and perceptual qualities such as tranquillity and naturalness that mapped data alone cannot capture. Finally, the classification and description stage (Step 4) distinguishes landscape character types (generic, recurring units sharing similar physical and cultural combinations) from landscape character areas (unique geographical entities with their own identity), maps their boundaries, and identifies key characteristics whose loss would fundamentally alter the character of a place (Tudor 2014). Across these steps, classification integrates objective spatial criteria with subjective interpretation (Simensen, Halvorsen, and Erikstad 2018; Terkenli, Gkoltsiou, and Kavroudakis 2021). For highways specifically, this expert procedure has been applied to corridor settings – for example, in characterising and rating the landscape of Malaysia's North-South Expressway from the road user's viewpoint (Jaal, Abdullah, and Ismail 2013). However, because such classifications rest on expert judgement applied to transitional, gradient-like landscapes, the resulting boundaries are rarely precise and the consistency of category assignment is seldom quantified.

Highway landscape classification adapts LCA principles to transportation contexts, addressing the distinctive characteristics of linear infrastructure corridors. Existing approaches can be grouped into several overlapping strands, each with distinct strengths and limitations.

First, function- and context-sensitive road environment classifications extend conventional road hierarchies by incorporating enriched corridor contexts. Expanded context classification differentiates five corridor characters – rural, rural town, suburban, urban, and urban core – based on land use, density, and building setbacks, linking them to context-sensitive, multimodal highway design expectations (Stamatiadis, Kirk, and Wright 2023). Vision-based classification similarly employs forward-facing imagery and machine-learning classifiers to distinguish between off-road and on-road scenes and to disaggregate urban streets, trunk roads, and highways (Tang and Breckon 2011). However, these frameworks prioritise transportation functionality (mobility, access, network connectivity) and operational classification over landscape perceptual quality and aesthetic character. Neither framework incorporates systematic, expert-based visual assessment protocols or quantitative landscape-preference metrics, nor do they address inter-rater reliability in defining landscape categories across diverse settings.

Second, land-use- and landform-based typologies group roadside scenes into recurring landscape characters, often to link them to public preferences or landscape ecological risk. (Kent 1993; Jaal et al. 2013; Gao et al. 2023; Merry et al. 2019) classify rural road or expressway views into land-use/land-cover categories (e.g., agriculture, forest, water, villages, industrial areas) and relate these characteristics to scenic quality, while Hu, He, and Xu (2023) and Zhang, Hu, et al. (2023) combine natural and anthropogenic factors to delineate themed sections and gradients in landscape ecological risk. However, these typologies are typically cross-sectional and viewpoint-based, derived from sampled photos or coarse segments, and rarely address classification reliability or inter-rater agreement, thereby limiting confidence in the consistency and reproducibility of landscape typologies.

Third, spatial-configuration-based classifications characterise highway visual environments using enclosure and elevation parameters, employing D/H ratios to distinguish enclosed, semi-open, and open configurations, and differentiating mountain, plain, and elevated sections (Guo et al. 2019; Qin et al. 2023). However, these approaches remain primarily confined to simulator experiments and are not well integrated

with character-based mapping or operational management metrics.

Fourth, element- and index-based visual-quality frameworks synthesise semantic segmentation or attribute-based scoring with perceptual ratings or spatial-quality indices to stratify views into visual-quality classes (Martín et al. 2018; Qin et al. 2024; Qin, Yang, and Wangari 2024). These frameworks capture fine-grained relationships between visual composition and perceived quality at specific locations, but typically presume a single authoritative classification per scene, give limited attention to inter-rater disagreement, and are rarely operationalised into reproducible, validated classification protocols with explicit uncertainty quantification. A smaller strand conceptualises roadside landscapes as cultural heritage assets, differentiating whether the road, the surrounding landscape matrix, or specific features constitute the primary object of conservation (Grazuleviciute-Vileniske and Matijosaitiene 2010); however, this perspective remains weakly connected to routine maintenance and safety protocols.

Despite methodological advances, highway landscape classification remains fragmented, emphasising transport functionality, static typologies, or localised indices. From a methodological standpoint, two limitations emerge across these approaches. First, classifications are typically treated as deterministic, with a single "correct" label per scene and little guidance on how to represent or propagate uncertainty and disagreement across raters or models. Second, although AI and computer vision approaches can automate large-scale image interpretation, they are rarely embedded in expert-anchored workflows that validate AI outputs and quantify agreement between AI and humans. In practice, these limitations manifest as persistent disagreements – for instance, distinguishing between palm plantations and mixed agricultural mosaics, or between low mountains and hilly terrain – even among experienced raters, undermining confidence in classification-based management decisions. These limitations point to the need for highway landscape classification frameworks that are explicitly uncertainty-aware, capable of quantifying classification reliability, and designed to integrate expert knowledge with AI-generated labels in a reproducible manner.

1.2. Recent AI advances in visual understanding: addressing classification challenges

The limitations identified above – especially the prevalence of deterministic single-label classifications and the absence of validated, expert-anchored frameworks – create opportunities for artificial intelligence. Traditional approaches are constrained by manual processing: they are labour-intensive, difficult to scale, and often opaque in translating subjective judgements into formal classes.

Recent advances in AI, particularly the development of multimodal large language models (MLLMs), offer transformative potential in addressing these challenges (Kaul, Enslin, and Gross 2020; Tan et al. 2023). Unlike earlier computer vision systems restricted to predefined categories, contemporary MLLMs can extract complex spatial and semantic patterns, identify subtle landscape characteristics, and scale analysis across extensive highway networks without proportional increases in human resources. Leading contemporary examples include ChatGPT-5.4 Thinking (OpenAI 2026), Claude Sonnet 4.6 (Anthropic 2026), and Gemini 3 Thinking (Google 2025), which demonstrate strong multimodal reasoning and visual understanding capabilities relevant to landscape interpretation.

Crucially, these models transcend the constraints of single-criterion classification by employing multimodal reasoning. They can simultaneously process multiple data types – street-view photographs, topographic data, and textual descriptions – within unified frameworks (Koyun and Taskent 2025; Varga 2025). This integration directly addresses the fragmentation in current approaches, where visual, ecological, and cultural dimensions are often assessed separately rather than holistically.

Furthermore, AI classification methods offer critical advances that directly address the two limitations identified above. First, when multiple models or repeated prompts are employed, AI systems can generate distributional outputs rather than single deterministic labels, providing a basis for quantifying classification uncertainty and inter-model disagreement. Second, the scalability and consistency of AI classification across large image datasets create opportunities for systematic comparison with expert annotations, enabling transparent validation of AI reliability and the identification of landscape features where automated classification can supplement or replace manual assessment.

However, these capabilities do not eliminate the need for human expertise. Without explicit anchoring to expert-defined taxonomies, AI outputs risk conceptual drift, inconsistent category boundaries, and opaque reasoning. The challenge is to design expert-anchored workflows that validate AI reliability while quantifying classification uncertainty. This integration remains underexplored in highway landscape research and is the focus of this study.

1.3. Research problems and objectives

Taken together, the preceding review reveals a fundamental methodological gap: existing highway landscape classification approaches fail to

systematically quantify classification reliability or validate AI-driven automation against expert judgment. Addressing these two shortcomings requires frameworks that can jointly evaluate human and AI classifications on shared datasets and explicitly characterise patterns of agreement and disagreement. This study addresses both challenges by developing a triangulated AI – expert validation framework that integrates human expertise with three leading multimodal large language models (ChatGPT-5.4 Thinking, Claude Sonnet 4.6, and Gemini 3 Thinking) to classify 16 highway landscape characters across 1,828 images. The specific objectives are:

(1) To establish a triangulated validation framework that compares expert annotations with AI-generated zero-shot classifications across 16 highway landscape characters;
(2) To evaluate and compare the performance of ChatGPT-5.4 Thinking, Claude Sonnet 4.6, and Gemini 3 Thinking in highway landscape classification, identifying patterns of AI-human agreement and divergence;
(3) To determine which landscape characters achieve reliable consensus versus persistent disagreement between AI systems and expert judgment;

Collectively, these objectives position the study at the intersection of corridor-scale visual assessment, expert judgement and multimodal AI, and motivate the empirical analyses that follow.

In summary, this study makes three key contributions. First, it operationalises an uncertainty-aware, AI-assisted framework treating class labels as testable outcomes. Second, it empirically compares ChatGPT-5.4 Thinking, Claude Sonnet 4.6, and Gemini 3 Thinking across 1,828 images, identifying where automated classifications are reliable and where they diverge. Third, it produces a stratified, quality-controlled dataset of 1,352 validated highway landscape images and an agreement-based reliability scheme that can directly support corridor-scale diagnostic mapping, risk-prioritised field verification, and the fine-tuning of future region-specific models.

2. Materials and methods

As shown in Figure 1, the conceptual framework of this study comprises four components: (1) study corridor selection, focusing on a 418-km segment of the Northern Route (E1) of Malaysia's North-South Expressway; (2) multi-source image acquisition, yielding 1,828 roadside images captured at approximately 250 m intervals using vehicle-mounted cameras and Google Street View; (3) expert-based landscape character classification using a three-dimensional taxonomy (landform, land use, and land cover) to derive 16 generalised highway landscape characters; and (4) triangulated AI – expert validation, in which zero-shot classifications from ChatGPT-5.4 Thinking, Claude Sonnet 4.6, and Gemini 3 Thinking are compared with expert labels and expert filtering retains 1,352 high-confidence samples for subsequent analysis.

2.1. Study area

The North-South Expressway (NSE) is Malaysia's primary transportation infrastructure, spanning 772 kilometres along the western coast of Peninsular Malaysia, from Bukit Kayu Hitam to Johor Bahru. It serves as a critical economic conduit for trade and tourism. The expressway's topographic profile varies substantially, characterised by predominantly level terrain in the northern and southern states, contrasted with rugged, hilly terrain through the central regions, particularly along the Titiwangsa Mountain Range (Jaal, Abdullah, and Ismail 2013). The NSE comprises two primary segments: the Northern Route (E1) extending from Bukit Kayu Hitam to Kuala Lumpur, and the Southern Route (E2) from Kuala Lumpur to Johor Bahru.

This research focuses on a strategically selected 418-kilometre section of the Northern Route (E1), extending from Plaza Tol Jalan Duta Arah Utara in Kuala Lumpur northward to Alor Setar Utara Toll Plaza in Kedah. This corridor traverses five key administrative regions – Kedah, Penang, Perak, Selangor, and the Federal Territory of Kuala Lumpur – encompassing diverse land-use patterns constituting a representative microcosm of Malaysia's economic and ecological landscape. To ensure the study covers the core categories of highway landscapes, the selected route integrates distinct visual environments: the extensive paddy cultivation systems in Kedah (representing open agricultural plains), the dense rubber and oil palm plantation zones in Perak and Selangor (representing semi-enclosed industrial vegetation), and the primary tropical forest ecosystems along rugged terrains (representing mountainous landscapes). These provide essential ecological functions, including carbon sequestration and biodiversity conservation.

2.2. Methods of the study

2.2.1. Collection of image data

The image dataset was constructed through a three-stage process: (1) field-based spatial referencing and sampling design, (2) high-resolution Google Street View (GSV) image acquisition, and (3) quality control and standardisation. This hybrid approach combines complementary sources: field imagery offers spatial control but variable quality; GSV imagery provides superior resolution but limited spatial precision. In

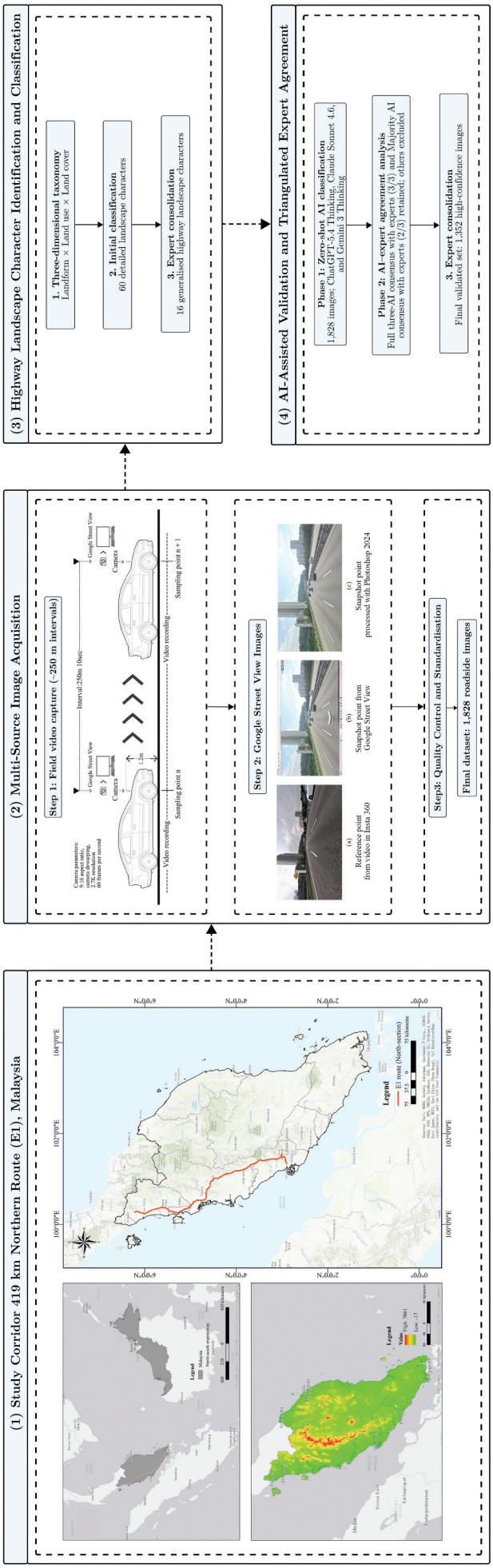


Figure 1. Research process for the expert-anchored, AI-assisted highway landscape classification and validation.

contrast, GSV imagery offers superior resolution and standardised viewing conditions but lacks precise spatial control over capture positions and orientations.

Step 1: Field-based Spatial Referencing and Sampling Design

Field data were collected using an Insta360 ONE X4 action camera with a 9:16 image aspect ratio, lens distortion correction enabled, 2.7K resolution, and 60 frames per second frame rate. The camera was mounted 1.2 metres above ground level on the vehicle's hood, aligning with the median driver's eye level of a standard passenger vehicle (Qi et al. 2025). Vehicle speed was maintained at approximately 90 km/h (below the 110 km/h legal maximum on Malaysian highways) for optimal image stability, with data collection occurring over two consecutive days (December 21–22, 2024), under clear weather conditions to minimise environmental variability.

Drivers' fixation points – their gaze locations – vary according to the type of surrounding landscape (Guo et al. 2019). Recognising the influence of highway landscapes on driver attention and emergency responsiveness, this study adopted a methodologically rigorous approach to characterise landscape variation along the corridor. A fixed 10-second sampling interval was adopted, corresponding to a spatial resolution of approximately 250 metres at a target speed of 90 km/h. This sampling density addresses a key limitation observed in previous highway landscape studies, which adopted wider intervals, such as 4-kilometre spacing (Jaal, Abdullah, and Ismail 2013) or 1-kilometre spacing (Bai, Ji, and Qi 2024). Such coarse samples may miss abrupt landscape transitions, particularly in segments with mixed land uses or varying topography. The 418-kilometre corridor generated 5 hours and 10 minutes of footage, yielding 1,860 spatial reference points. This exceeded the theoretical estimate of 1,676 points due to traffic interactions and topographical speed variations.

Step 2: High-Resolution GSV Image Acquisition

GSV imagery was selected as the primary data source for landscape analysis due to its high-resolution imagery, consistent viewpoint, and extensive spatial coverage (Kelly et al. 2012). Despite limitations including incomplete capture metadata (L. Yin et al. 2015) and temporal inconsistencies from infrequent updates (P. Jiang et al. 2024), GSV was deemed suitable given the study's focus on landscape composition rather than temporal dynamics.

Spatial correspondence between field reference points and GSV locations was verified through systematic comparison using Google Earth Pro. To ensure comparability across data sources, each GSV image was captured at the corresponding field-derived sampling point (Step 1), enabling spatial alignment between field and GSV datasets. High-resolution images were obtained using standardised full-screen capture protocols. Post-processing was conducted in Adobe Photoshop (Version 2024), including precise cropping to remove interface elements using content-aware fill while preserving landscape fidelity. This procedure yielded 1,860 GSV images that were spatially aligned with the field-based image dataset. It should be emphasised that the field video footage and Google Earth Pro were used solely for sampling design and for cross-verifying capture locations; no satellite imagery was used in the analysis, and the field and GSV streams were not merged. Only the resulting GSV images served as the visual input to the classification stage, while the field-based points provided the spatial reference against which each GSV view was aligned.

Step 3: Quality Control and Standardisation

Quality screening excluded images compromised by environmental obstructions (such as fog, haze, or low visibility) or foreground vehicles that obscured landscape elements. Following screening, 1,828 high-resolution images were retained (98.3% retention rate) and systematically renamed using a sequential numerical format for analytical compatibility.

2.2.2. Highway landscape character identification and classification

Highway landscapes function as linear visual corridors influenced by spatial openness, landform complexity, vegetation density, and built elements. Classification requires standardised criteria across the dimensions of landforms, land uses, and land covers (Gao et al. 2024).

This study employs a three-dimensional classification comprising:

- landform (e.g., mountains, hills, and associated geomorphic characters)
- land use (e.g., agricultural, industrial, and urban zones), and
- land cover (e.g., vegetation, built structures, and other surface elements).

In the first step, trained researchers applied the three-dimensional framework (landform – land use – land cover) to inductively describe the dominant visual character of the corridor scene. This open coding process produced 59 fine-grained corridor characters (e.g., "highway corridor through cut slopes with dominant teak trees and mountain backdrop", "highway corridor through agricultural plains with roadside tree plantings", "highway corridor through extensive palm plantations with mountain backdrop").

The preliminary classification was reviewed and consolidated by two authors with domain expertise in highway landscape and corridor planning. Expert eligibility was defined by the following criteria: (i) possession of a postgraduate degree in landscape architecture, urban planning or a related discipline; (ii) substantial professional or research experience in highway landscape

design, visual impact assessment or corridor planning; and (iii) demonstrated familiarity with the Malaysian highway context through prior project involvement or published research. The two experts then independently reviewed the 59 preliminary characters together with representative images from each group.

They noted that several categories were distinguished only by subtle contextual differences, and that some characters contained very few samples, which would undermine subsequent statistical analysis and model evaluation. To address these issues, the experts conducted an iterative consolidation process guided by three principles: (i) each final character had to represent a clear dominant; (ii) class labels had to remain visually recognisable and conceptually generic (e.g., "highway corridor with dense vegetation" rather than site-specific descriptions); and (iii) the resulting taxonomy had to avoid minimal classes while preserving meaningful variation in corridor characters.

Following several rounds of discussion, the experts reviewed each of the 59 preliminary characters individually and assigned them to final characters based on the dominant visual elements in representative images. This step recognised that an individual character might be better characterised by visual features different from those emphasised in its initial description. For instance, a character initially described in terms of slope and vegetation might be reassigned to a structure-related character if built elements (e.g., advertisements or overpasses) visually dominated the scene; similarly, a character initially described in terms of a palm-oil plantation might be allocated to "highway corridor with utility" if electrical infrastructure constituted the primary visual element rather than plantation vegetation.

Through this image-based review process, the experts identified 16 generalised highway landscape characters that captured the full range of dominant visual conditions along the corridor (Table 2). These final characters can be broadly grouped into four thematic clusters based on their primary visual attributes: (i) vegetation-dominated characters, including "highway corridor with dense vegetation", "highway corridor with light vegetation", "highway corridor with dense palm plantation", and "highway corridor with paddy and vegetation"; (ii) slope – vegetation characters, including "highway corridor with slope dense vegetation", "highway corridor with slope light vegetation" and "highway corridor with rocky cliff"; (iii) geomorphologically distinctive characters, including "highway corridor with limestone hill" and "highway corridor with mountain"; and (iv) infrastructure- and structure-related characters, including "highway corridor with urban setting", "highway corridor with development view", "highway corridor with advertisement", "highway corridor with slope and advertisement", "highway corridor with utility", "highway corridor with bridge" and "highway corridor with engineered wall".

In this taxonomy, each character label denotes the visually dominant element or combination of elements in the corridor scene. For example, "highway corridor with dense vegetation" refers to scenes where continuous tree or shrub cover visually dominates both sides or the horizon of the corridor, whereas "highway corridor with bridge" refers to scenes in which an overpass or viaduct is the primary visual feature in the driver's field of view. It is therefore best understood as an interpretive but transparent classification scheme that standardises how experts describe corridor scenes, rather than as a uniquely "correct" segmentation of all possible highway landscapes.

2.2.3. AI-assisted validation and triangulated agreement

To ensure the robustness and interpretability of the highway landscape classification scheme, this study implemented a structured, triangulated validation approach: (1) zero-shot AI-assisted classification, (2) multi-rater agreement analysis incorporating expert and AI perspectives, and (3) expert-led filtering of reliable image samples.

2.2.3.1 AI-assisted classification validation using ChatGPT- 5.4 Thinking, Claude Sonnet 4.6, and Gemini 3 Thinking. 1. AI-Assisted Classification Validation Using ChatGPT-5.4 Thinking, Claude Sonnet 4.6, and Gemini 3 Thinking

Each of the 1,828 highway landscape images was independently classified using three state-of-the-art multimodal large language models (MLLMs) – ChatGPT-5.4 Thinking, Claude Sonnet 4.6, and Gemini 3 Thinking – via zero-shot classification. In other words, the models were not fine-tuned on any of our specific images or characters beforehand; instead, they were prompted to assign one of the 16 landscape character labels to each image based solely on its visual content. Both AI services were accessed via paid subscriptions to ensure full functionality and consistent batch processing protocols.

Images were uploaded one by one to the AI models, and only the first character label output from each model was recorded (no iterative feedback or additional context, like GPS coordinates, was provided). A standardised prompt was used to enforce consistency in how the models understood the task and the available characters. The prompt to the AI was structured as follows:

> I have collected a series of highway landscape images that require classification. I will upload these images one by one for your review. Please identify which landscape character each belongs to from the following 16 predefined options: Highway corridor with advertisement, Highway corridor with bridge, Highway corridor with dense palm plantation, Highway corridor with dense vegetation, Highway corridor with development view, Highway corridor with engineered wall, Highway

corridor with light vegetation, Highway corridor with limestone hill, Highway corridor with mountain, Highway corridor with paddy and vegetation, Highway corridor with rocky cliff, Highway corridor with slope and advertisement, Highway corridor with slope dense vegetation, Highway corridor with slope light vegetation, Highway corridor with urban setting, Highway corridor with utility.

This prompt ensured that all three AI models chose from the same set of character labels defined by our experts, facilitating a direct comparison with the human-classified labels. This prompt ensured that all three AI models chose from the same set of character labels defined by our experts, facilitating a direct comparison with the human-classified labels. It should be emphasised that no supervised CNN or ViT classifier was trained, fine-tuned, or applied at any stage of this study; the supervised pipeline is invoked here only as a conceptual comparison to explain why a zero-shot MLLM design was deliberately preferred. Although the classification task could, in principle, be addressed with a conventional supervised pipeline (e.g., training a task-specific convolutional neural network or vision transformer on the expert-labelled images), we deliberately adopted a zero-shot MLLM approach for several methodological and practical reasons.

First, in this study, AI is positioned as a validation tool rather than as the primary classifier: the objective is not to replace expert judgement with a maximally accurate model, but to quantify classification reliability and highlight ambiguous cases that warrant closer expert review. In this context, the observed 67–70% agreement between zero-shot AI and the expert provides diagnostic information about which landscape characters are conceptually stable versus inherently ambiguous – a diagnostic capability that high-accuracy but opaque supervised models trained on the same labels cannot provide – rather than serving solely as a performance score.

Second, training and evaluating a supervised CNN/ViT on the same limited set of expert-labelled images would primarily measure how well the model reproduces the expert's taxonomy, potentially conflating model performance with taxonomic validity (Geiger et al. 2021; Northcutt, Jiang, and Chuang 2021). By contrast, zero-shot MLLMs are pre-trained on broad visual – semantic corpora and are used here with fixed parameters, providing an independent source of judgements that can reveal where expert-defined categories align with, or deviate from, more general visual semantics (Radford et al. 2021; Zhai et al. 2022).

Third, the zero-shot setting eliminates the need for task-specific model training, hyperparameter tuning, and GPU-intensive optimisation at the application stage, making the protocol more easily reproducible and implementable by highway agencies (Novack et al. 2023; Radford et al. 2021). Finally, modern multimodal LLMs provide human-readable rationales for their decisions, allowing experts to inspect why particular classifications or disagreements occur, whereas typical CNN/ViT outputs often require additional post-hoc analysis to be interpreted (Kazmierczak et al. 2025; K. Li, Vosselman, and Yang 2025).

As an example of the AI output, Table 1 presents an excerpt of classification results for one sample image, including the character chosen by each model and its justifications. In this example, all three models correctly identified the scene as "Highway corridor with paddy and vegetation," demonstrating the methodology's effectiveness.

2.2.3.2. Triangulated agreement framework. 2. Triangulated Agreement Framework

To ensure the reliability and interpretability of landscape classification outcomes, this study introduces a triangulated agreement framework that simultaneously considers expert annotations, ChatGPT-5.4 Thinking predictions, Claude Sonnet 4.6 predictions, and Gemini 3 Thinking predictions. This framework extends beyond traditional AI-human comparison by incorporating a multi-model triangulated validation structure, thereby offering a more nuanced understanding of classification uncertainty and inter-model concordance.

The framework combines four independent sources of judgement-one expert baseline and three MLLMs – to map a graded agreement structure. It is therefore not an ensemble in the classical machine-learning sense of aggregating models to optimise predictive accuracy, but a triangulated comparison design that treats disagreement patterns themselves as informative signals about semantic stability and boundary fragility in the taxonomy.

Based on the number of AI models agreeing with the expert annotation, each image was assigned to one of seven agreement categories, organised into four hierarchical levels:

- Level 1: Full consensus (3/3 AI = Expert): (A) All three AI models agree with the expert.
- Level 2: Majority consensus (2/3 AI = Expert): (B) Two AI models agree with the expert, and one disagrees.
- Level 3: Minority consensus (1/3 AI = Expert): (C) One AI model agrees with the expert, while the other two provide different non-expert labels. (D) One AI model agrees with the expert, whereas the other two AI models agree with each other on a different label.
- Level 4: No consensus (0/3 AI = Expert): (E) None of the three AI models agrees with the expert, and all three provide different labels. (F) None of the three AI models agrees with the expert; two agree with each other, while the third assigns a different label. (G) All three AI models agree with each other, but all disagree with the expert.

Table 1. Example of classification results and answer generated by ChatGPT-5.4 Thinking, Claude Sonnet 4.6, and Gemini 3 Thinking for highway landscape images.

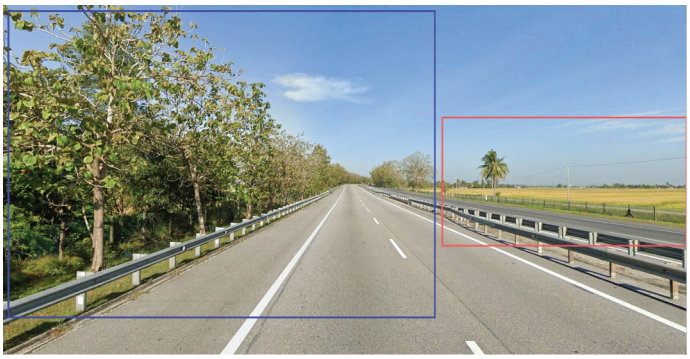
ChatGPT-5.4 Thinking	 This image should be classified as: Highway corridor with paddy and vegetation Reason: • The right side shows a broad, open agricultural field that appears to be paddy land, while the left side features a continuous line of roadside trees/vegetation. There are no dominant features such as bridges, steep slopes, limestone hills, dense palm plantations, urban elements, or advertisements.
Claude Sonnet 4.6	Highway corridor with paddy and vegetation This image shows a clear highway scene featuring: ● Right side: Expansive golden/yellow paddy (rice) fields stretching to the horizon, characteristic of northern Malaysia's agricultural plains (consistent with the Ipoh corridor area) ● Left side: A tree-lined roadside with mixed deciduous vegetation, including some dry/seasonally bare trees ● Centre: A dual-carriageway expressway with a central crash barrier dividing the lanes ● Background: Open flat landscape with scattered coconut palms and low-rise structures, under a clear blue sky The dominant landscape character is the juxtaposition of the highway infrastructure against active paddy cultivation on one side and roadside vegetation on the other, which aligns precisely with the "Highway corridor with paddy and vegetation" typology in your 16-class taxonomy.
Gemini 3 Thinking	Based on the visual elements in the image, here is the classification for this landscape: Classification Highway corridor with paddy and vegetation Reasoning: ● Paddy Field: To the right of the highway, there is an expansive, flat, yellowish-green field that is characteristic of a paddy field (common along the North-South Expressway, particularly in northern states like Kedah or Perak). ● Vegetation: The left side of the corridor is lined with a row of mature trees, while a single palm tree stands out on the right, providing the "vegetation" component of this category. ● Clear Horizon: Unlike the "dense vegetation" or "mountain" categories, this scene features a low, open horizon dominated by agricultural land.

Table 2. Distribution and retention of Highway landscape characters following expert – AI triangulated validation.

No.	Name of Landscape Characters	Original Sample Counts	Original Percentage (%)	Validated Samples	Retention Rate (%)
1	Highway corridor with advertisement	132	7.2%	120	90.9%
2	Highway corridor with bridge	103	5.6%	91	88.3%
3	Highway corridor with dense palm plantation	160	8.8%	139	86.9%
4	Highway corridor with dense vegetation	289	15.8%	185	64.0%
5	Highway corridor with development view	177	9.7%	127	71.8%
6	Highway corridor with engineered wall	48	2.6%	29	60.4%
7	Highway corridor with light vegetation	148	8.1%	84	56.8%
8	Highway corridor with limestone hill	60	3.3%	55	91.7%
9	Highway corridor with mountain	58	3.2%	48	82.8%
10	Highway corridor with paddy and vegetation	71	3.9%	55	77.5%
11	Highway corridor with rocky cliff	32	1.8%	32	100.0%
12	Highway corridor with slope and advertisement	70	3.8%	63	90.0%
13	Highway corridor with slope dense vegetation	198	10.8%	101	51.0%
14	Highway corridor with slope light vegetation	169	9.2%	147	87.0%
15	Highway corridor with urban setting	23	1.3%	13	56.5%
16	Highway corridor with utility	90	4.9%	63	70.0%
Total		1828	100%	1352	74.0%

A total of 1,828 highway landscape images were analysed under this framework. The distribution of agreement categories (Figure 2) revealed that 837 images (45.8%) achieved full consensus (Category A), while 515 (28.2%) reached majority consensus (Category B). Together, these two levels yielded 1,352 high-confidence images (74.0%) retained for subsequent analysis. Among the excluded images, minority consensus (Level 3) accounted for 226 cases (12.4%). Within this level, Category C – where one AI model agreed with the expert while the other two provided different non-expert labels – comprised 85 images (4.7%), and Category D – where one AI model agreed with the expert while the other two converged on an alternative label – comprised 141 images (7.7%). No consensus (Level 4) accounted for 250 cases (13.6%).

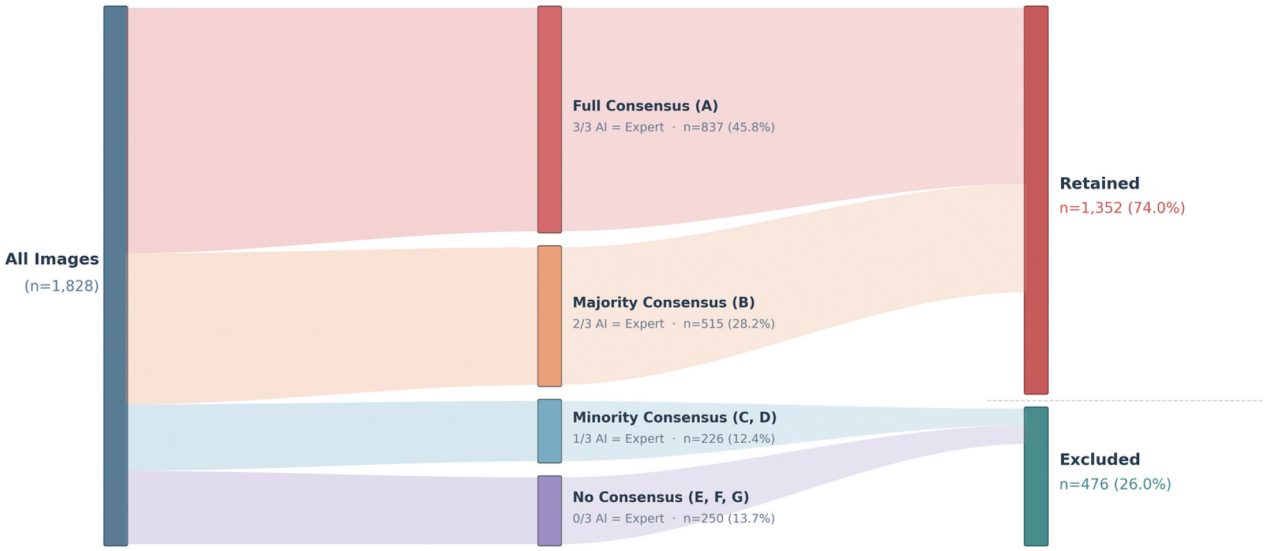


Figure 2. Distribution of agreement categories under the triangulated agreement framework.

Category E, in which all three AI models and the expert each assigned different labels, comprised 22 images (1.2%). Category F, in which two AI models agreed with each other while the third gave another non-expert label, comprised 112 images (6.1%). Category G, in which all three AI models converged on the same label yet diverged from the expert, accounted for 116 images (6.3%), suggesting potential ambiguities in either the expert annotation or the taxonomic boundary definition for those landscapes. Detailed classification information for the three models is provided in Supplementary Table S1.

2.2.3.3 Expert Review and Sample Retention.

Following the triangulated agreement analysis, expert reviewers selected high-confidence samples based on the concordance levels defined above. Images were retained only if they achieved full three-AI consensus with experts (Category A: all three AI models agreed with the expert classification) or majority AI consensus with experts (Category B: two of the three AI models agreed with the expert). Images falling below the majority consensus threshold – including minority consensus (Categories C and D) and no consensus (Categories E, F, and G) – were excluded from subsequent analysis. This filtering process ensured that only images with robust multi-source validation progressed to the next analysis stage. Of the original 1,828 images, 1,352 (74.0%) met the retention criteria (Table 2).

2.2.4. Evaluation metrics and confusion matrix analysis

Standard character performance metrics were employed to evaluate the AI models' classification performance (ChatGPT-5.4 Thinking, Claude Sonnet 4.6, and Gemini 3 Thinking), including accuracy, precision, recall, and F1 Score. These metrics assess model behaviour across both balanced and imbalanced classes, capturing complementary dimensions of classification performance.

All metrics were computed based on the confusion matrix – a square matrix where each row represents the predicted classes and each column corresponds to the actual (expert-labelled) classes. Diagonal entries represent correct classifications, while off-diagonal entries indicate misclassifications.

The metrics used for performance evaluation are defined and calculated using the following equations:

- Accuracy: Proportion of correctly classified images relative to the total dataset.

$$\text{Accuracy} = \frac{\sum_{i=1}^{N} \text{TP}_i}{\text{Total number of samples}} \tag{1}$$

Where: N is the number of classes in the classification task; TPi is the number of true positives for class i, i.e., the number of samples correctly predicted as belonging to class i; the denominator, "Total number of samples," refers to all samples across all classes in the dataset.

- Precision: Proportion of images correctly assigned to a given character relative to all images predicted as that character.

$$\text{Precision}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i} \tag{2}$$

Where: TP *i* (True Positives) is the number of images correctly classified as class *i*; FP *i* (False Positives) is the number of images incorrectly classified as class *i* (i.e., predicted as class *i*, but belong to other classes); *i* denotes a specific landscape character or class.

- Recall: Proportion of correctly identified images relative to all actual images in a character.

$$\text{Recall}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \tag{3}$$

Where: TP *i* (True Positives) is the number of images correctly classified as class *i*; FN *i* (False Negatives) is the number of images that belong to class *i* but were incorrectly predicted as other classes; *i* denotes a specific landscape character or class.

- F1-Score: The harmonic mean of precision and recall provides a balanced measure of the model's ability to avoid false positives and negatives.

$$\text{F1}_i = 2 \times \frac{\text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{4}$$

Where Precision_i is the precision for class *i*; Recalli is the recall for class *i*; *i* refers to a specific landscape character or class.

Additionally, macro-averaged and weighted F1-scores were calculated to summarise performance across all classes, accounting for class balance and variability.

- Macro-averaged F1-score: The unweighted average of F1-scores across all classes.

$$\text{Macro\_F1} = \frac{1}{N}\sum_{i=1}^{N} \text{F1}_i \tag{5}$$

Where: *N* is the total number of classes; F1_i is the F1-score for class *i*.

- Weighted F1-scores: The average F1-score weighted by the number of true samples in each class.

$$\text{Weighted} - \text{F1} = \sum_{i=1}^{N} \frac{n_i}{\sum_{j=1}^{N} n_j} \cdot \text{F1}_i \tag{6}$$

Where: *N* is the total number of classes;n_i is the number of true samples belonging to class *i*; F1_i is the F1-score for class *i*; The term $\frac{n_i}{\sum_{j=1}^{N} n_j}$ represents the proportion of true samples in class iii relative to the total number of samples across all classes.

Note: In landscape classification tasks with class imbalance – such as the present study – F1-score is particularly important because it balances both false positives and false negatives, offering a more stable metric than accuracy alone.

To assess performance differences among the three AI models, a two-level statistical testing strategy was adopted. At the overall accuracy level, Cochran's Q test was first applied as an omnibus test to evaluate whether classification accuracy differed significantly across the three models. Where the omnibus test was significant, pairwise McNemar's tests with Bonferroni correction (adjusted $\alpha = 0.05/3 = 0.0167$) were conducted to identify which specific model pairs differed. At the per-class F1 level, the Friedman test was applied as a non-parametric omnibus test for three related samples, followed by pairwise Wilcoxon signed-rank tests with Bonferroni correction. The Wilcoxon test was chosen because, although the Shapiro–Wilk tests on F1-score differences indicated approximate normality ($p > 0.05$ for all three pairs), the small sample size ($n = 16$ classes) favoured a non-parametric approach to avoid distributional assumptions. Additionally, a 10,000-iteration non-parametric bootstrap was used to obtain 95% confidence intervals for the mean pairwise F1-score gaps, providing a distribution-free estimate of the magnitude and uncertainty of performance differences.

2.2.5. Cohen's Kappa statistic

Cohen's Kappa (κ) was employed to assess agreement between AI predictions and expert annotations beyond chance expectation. Unlike accuracy, Cohen's Kappa accounts for the probability of random agreement, providing a more robust measure of classification reliability.

The Kappa value is calculated as follows:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{7}$$

Where: *Po* is the observed agreement, i.e., the proportion of samples where the model prediction matches the expert label; *Pe* is the expected agreement by chance, calculated from the marginal totals of the confusion matrix as:

$$P_e = \sum_{i=1}^{N} \left(\frac{A_i}{N} \times \frac{B_i}{N} \right) \tag{8}$$

Where: *Ai* is the number of samples predicted in class *i* by the model, *Bi* is the number of samples labelled as class *i* by the expert, and *N* is the total number of samples. Kappa values were interpreted according to the widely adopted scale proposed by Landis and Koch (1977), where values greater than 0.80 indicate almost perfect agreement, 0.61–0.80 substantial agreement, 0.41–0.60 moderate agreement, 0.21–0.40 fair agreement, and values below 0.20 suggest slight or poor agreement.

2.2.6. Supporting tools

This study utilised Python 3.11 (PyCharm 2024.1) for data processing, with key libraries including Pandas, GeoPandas, scikit-learn for data handling, and Matplotlib for visualisation. Google Earth Pro was

used for geospatial referencing and verification, Adobe Photoshop 2024 for image enhancement and standardisation, and Lucidchart for creating a flowchart that summarised all stages of data collection leading to AI-assisted expert validation.

3. Results

3.1. Image distribution and validation outcomes across highway landscape characters

Addressing Objective 1, this subsection summarises how the expert – AI workflow filters and stratifies the 16 landscape characters. The dataset comprised 1,828 highway landscape images, predominantly from the northern section of the corridor, distributed across 16 characters (Table 2). Vegetation-dominated characters – led by "dense vegetation" (15.8%) and "slope dense vegetation" (10.8%) – accounted for 56.6% of observations, reflecting the corridor's predominantly vegetated character, shaped by Malaysia's tropical climate. Infrastructure-related characters, including "development view," "advertisement," and "bridge," represented a further 33.8%, while geomorphological features ("limestone hill," "mountain," and "rocky cliff") comprised only 8.3%. "Urban setting" was the scarcest character (1.3%), indicating limited urbanised segments along the studied route.

The triangulated validation retained 1,352 images (74.0%), with retention rates stratified by morphological distinctiveness and semantic clarity. Characters with morphologically distinct or visually unambiguous features achieved high retention (≥90%), including "rocky cliff" (100.0%), "limestone hill" (91.7%), "advertisement" (90.9%), and "slope and advertisement" (90.0%). These high-consensus characters share two key attributes: well-defined visual boundaries with minimal transitional overlap, and distinctive visual signatures that enable consistent human – AI interpretation.

Moderately complex characters showed intermediate retention (70–89%), ranging from "bridge" (88.3%) to "utility" (70.0%). Classification ambiguity in this tier arose from contextual variability rather than fundamental semantic confusion. For instance, "mountain" scenes were sometimes conflated with slope – vegetation characters when foreground vegetation partially obscured the mountain backdrop, while "dense palm plantation" was occasionally misclassified as other vegetation-dominated characters when the regular geometry of plantation rows was not clearly discernible.

Transitional or semantically ambiguous characters exhibited low retention (<70%), notably "slope dense vegetation" (51.0%), "urban setting" (56.5%), "light vegetation" (56.8%), "engineered wall" (60.4%), and "dense vegetation" (64.0%). Classification disagreement in these characters is structurally embedded, driven by two interrelated sources of ambiguity: first, vegetation density and the built-environment spectrum both exist along gradients where categorical boundaries are inherently subjective (e.g., "light" versus "dense" vegetation; "urban setting" versus "development view"); and second, composite scenes involving co-occurring elements – such as slope – vegetation combinations or utility infrastructure as a secondary feature – require simultaneous multi-attribute judgements in which the dominant character is context-dependent.

3.2. AI-human concordance and discordance in landscape classification

To further address Objective 3, the classification agreement matrix provides a complementary view of how experts and AI models converge or diverge for each character (Figure 3). Classification agreement patterns revealed distinct performance tiers based on visual complexity: morphologically distinct landscapes achieved strong consensus, while transitional or ambiguous characters showed substantial disagreement.

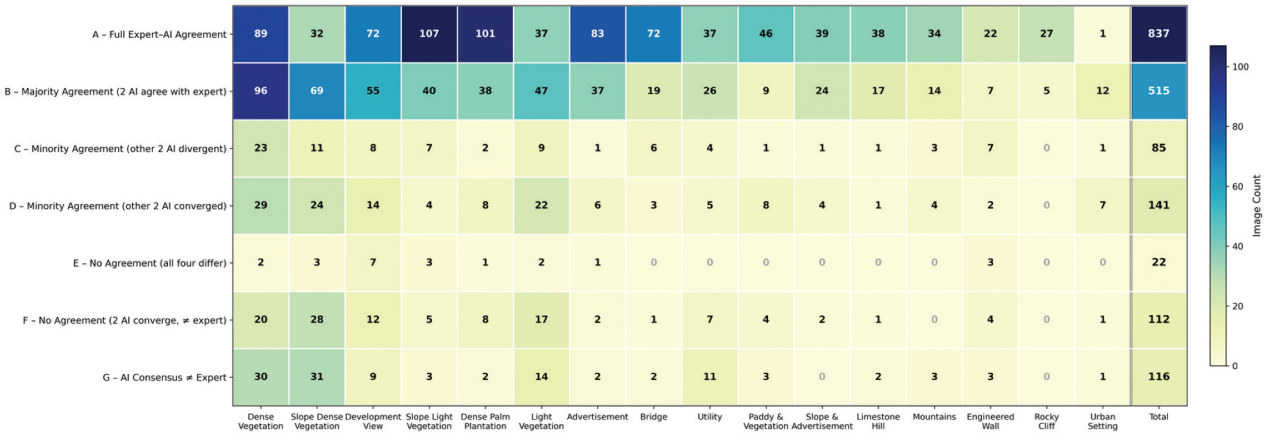


Figure 3. Distribution of classification agreement types across highway landscape characters based on expert and AI model comparison.

Among high-retention characters, several achieved Category A (full consensus) rates exceeding 60%, including "slope light vegetation" (63.3%), "dense palm plantation" (63.1%), and "advertisement" (62.9%). "Rocky cliff" was the most unambiguous character in the taxonomy, achieving full consensus in 84.4% of cases with zero instances in Categories C through G.

By contrast, complex built environments were considerably more challenging. "Urban setting" achieved only 1 full consensus case out of 23 samples, relying heavily on majority consensus (Category B: 12 cases) to reach its 56.5% retention rate. "Development view" and "engineered wall" exhibited similarly dispersed profiles across multiple agreement categories; notably, "engineered wall" had the highest proportion of Category C cases (7 of 48, 14.6%) relative to sample size, indicating highly fragmented classification signals.

A particularly informative diagnostic emerged from Category G, where all three AI models converged on the same label yet diverged from the expert. "Slope dense vegetation" (31 cases, 15.7%) and "dense vegetation" (30 cases, 10.4%) together accounted for over half of all Category G cases. This pattern suggests that the AI models detected coherent visual patterns that the expert taxonomy classified differently – signalling potential ambiguities in the taxonomic boundary between these two characters rather than purely model error. Category D (one AI agreed with the expert while the other two converged on an alternative) reinforced this finding, with "dense vegetation" (29 cases), "Slope dense vegetation" (24 cases), and "light vegetation" (22 cases) again dominating. Together, Categories D and G confirm that vegetation-density gradients represent the most structurally embedded source of taxonomic ambiguity, where competing classification interpretations systematically coexist.

These agreement patterns corroborate the three reliability tiers identified in Section 3.1 and directly address Objective 3 by pinpointing where automated classification is trustworthy and where semantic ambiguity necessitates expert oversight.

3.3. Comparative classification and confusion analysis of AI models and expert annotations

To address Objective 2, ChatGPT-5.4 Thinking, Claude Sonnet 4.6, and Gemini 3 Thinking were evaluated using multi-class accuracy, per-class precision, recall, and F1 metrics, statistical significance tests, and confusion-matrix-based error analysis. All three models independently classified all 1,828 images against expert annotations (Table 3; Figure 4).

3.3.1. Comparative classification performance of ChatGPT-5.4 Thinking, Claude Sonnet 4.6, and Gemini 3 Thinking

Claude Sonnet 4.6 demonstrated the strongest overall performance among the three models, achieving 70.3% accuracy and a 72.1% macro-averaged F1 score, followed by ChatGPT-5.4 Thinking (68.8%, 69.7%) and Gemini 3 Thinking (67.0%, 67.6%) (Table 3). Claude achieved the highest per-class F1 score in 9 of the 16 characters, ChatGPT in 4, and Gemini in 3, indicating that Claude generalises most robustly across diverse landscape types, though each model demonstrates distinct strengths.

Claude's performance advantages were most pronounced in infrastructure-related and transitional environments, substantially outperforming both counterparts for "engineered wall" (F1: 76.2% vs. 61.7% and 74.4%), "mountain" (70.1% vs. 60.1% and 60.7%), and "slope and advertisement" (87.8% vs. 76.6% and 77.4%). Claude also led in "development view," primarily through higher precision (80.1%), suggesting fewer false positives in heterogeneous built-up contexts.

ChatGPT, while ranking second overall, achieved the highest F1 scores for "urban setting" (54.2% vs. 44.8% and 35.0%) and "bridge" (80.2% vs. 74.7% and 75.0%), indicating greater sensitivity to built-environment attributes in these contexts. ChatGPT also had a marginal advantage in "light vegetation" and "slope light vegetation."

Gemini 3 Thinking, despite having the lowest overall accuracy, achieved the best F1 scores for morphologically distinctive characters, including "rocky cliff" (95.4%), "limestone hill" (87.0%), and "dense vegetation" (70.0%). These results suggest that Gemini's visual encoder is particularly responsive to high-contrast textures, though this advantage does not extend to semantically ambiguous characters.

Characters featuring subtle visual gradients or overlapping semantics (e.g., "light vegetation," "slope dense vegetation," "urban setting") remained challenging for all three systems, with F1 scores generally below 60%, underscoring structural difficulty rather than the limitations of any single model. Conversely, morphologically distinctive characters achieved high F1 scores (>73%) across all three models with small inter-model differences, confirming that visually apparent class boundaries yield only incremental gains from differences in model architecture. Overall, this pattern of distributed expertise – Claude in infrastructure-heavy and transitional environments, ChatGPT in specific built-environment characters, Gemini in geomorphologically distinctive features – strengthens the case for multi-model triangulation, as no single model consistently dominates across all landscape characters.

Table 3. Comparative precision, recall, and F1 scores of ChatGPT, Claude, and Gemini across 16 Highway landscape characters.

No.	Name of Landscape Characters	ChatGPT Precision (%)	Claude Precision (%)	Gemini Precision (%)	ChatGPT Recall (%)	Claude Recall (%)	Gemini Recall (%)	ChatGPT F1-score (%)	Claude F1-score (%)	Gemini F1-score (%)	support
1	Highway corridor with advertisement	61.4%	57.1%	56.6%	77.3%	91.7%	81.1%	68.5%	70.3%	66.7%	132
2	Highway corridor with bridge	70.9%	71.1%	67.4%	92.2%	78.6%	84.5%	80.2%	74.7%	75.0%	103
3	Highway corridor with dense palm plantation	66.5%	76.7%	69.3%	83.1%	82.5%	77.5%	73.9%	79.5%	73.2%	160
4	Highway corridor with dense vegetation	78.6%	86.9%	75.3%	58.5%	52.9%	65.4%	67.1%	65.8%	70.0%	289
5	Highway corridor with development view	69.2%	80.1%	71.2%	67.2%	63.8%	65.5%	68.2%	71.1%	68.2%	177
6	Highway corridor with engineered wall	75.8%	88.9%	84.2%	52.1%	66.7%	66.7%	61.7%	76.2%	74.4%	48
7	Highway corridor with light vegetation	60.6%	48.6%	63.7%	56.1%	68.9%	34.5%	58.2%	57.0%	44.7%	148
8	Highway corridor with limestone hill	88.9%	86.7%	90.9%	80.0%	86.7%	83.3%	84.2%	86.7%	87.0%	60
9	Highway corridor with mountain	48.4%	61.8%	50.6%	79.3%	81.0%	75.9%	60.1%	70.1%	60.7%	58
10	Highway corridor with paddy and vegetation	78.5%	77.3%	68.3%	71.8%	81.7%	78.9%	75.0%	79.5%	73.2%	71
11	Highway corridor with rocky cliff	96.7%	88.6%	93.9%	90.6%	96.9%	96.9%	93.5%	92.5%	95.4%	32
12	Highway corridor with slope and advertisement	84.5%	88.4%	70.6%	70.0%	87.1%	85.7%	76.6%	87.8%	77.4%	70
13	Highway corridor with slope dense vegetation	74.0%	77.5%	69.7%	46.0%	47.0%	42.9%	56.7%	58.5%	53.1%	198
14	Highway corridor with slope light vegetation	65.6%	64.4%	58.7%	84.6%	83.4%	75.7%	73.9%	72.7%	66.1%	169
15	Highway corridor with urban setting	52.0%	34.1%	41.2%	56.5%	65.2%	30.4%	54.2%	44.8%	35.0%	23
16	Highway corridor with utility	59.2%	76.8%	58.0%	67.8%	58.9%	64.4%	63.2%	66.7%	61.1%	90
	Accuracy	68.8%	70.3%	67.0%	68.8%	70.3%	67.0%	68.8%	70.3%	67.0%	1828
	Macro avg	70.7%	72.8%	68.1%	70.8%	74.6%	69.3%	69.7%	72.1%	67.6%	1828
	Weighted avg	70.3%	73.9%	68.1%	68.8%	70.3%	67.0%	68.4%	70.2%	66.3%	1828

3.3.2. Statistical significance of performance differences

To validate the observed performance differences, statistical significance testing was conducted at both the overall accuracy and per-class F1 levels, with Bonferroni correction applied to all pairwise comparisons (adjusted $\alpha = 0.05/3 = 0.0167$).

Cochran's Q test yielded a statistically significant result ($Q = 7.30$, $df = 2$, $p = 0.026$), confirming that the three models do not perform equivalently on the classification task. Pairwise McNemar's tests with Bonferroni correction identified the Claude – Gemini comparison as the only pair to reach significance ($\chi^2 = 6.90$, $p = 0.009$), corresponding to a 3.3 percentage-point accuracy advantage for Claude (70.3% vs 67.0%). The Claude – ChatGPT difference (1.5 percentage points; $\chi^2 = 1.67$, $p = 0.197$) and the ChatGPT – Gemini difference (1.8 percentage points; $\chi^2 = 2.09$, $p = 0.148$) were not statistically significant after correction.

At the per-class level, a Friedman test on the 16 paired F1 scores was not statistically significant ($\chi^2 = 3.88$, $p = 0.144$), indicating that the overall rank ordering of models did not differ consistently across all landscape characters. Pairwise Wilcoxon signed-rank tests confirmed the accuracy-level pattern: Claude's F1 scores significantly exceeded Gemini's ($W = 15.0$, $p = 0.004$; mean difference = +4.54 percentage points), while neither the Claude – ChatGPT ($W = 34.0$, $p = 0.083$) nor ChatGPT – Gemini ($W = 47.0$, $p = 0.298$) comparisons reached significance. Bootstrap resampling (10,000 iterations) corroborated these results, with only the Claude – Gemini 95% confidence interval excluding zero ([+2.27, +6.78] percentage points).

These results establish a performance hierarchy of Claude > ChatGPT >Gemini, but only the gap between Claude and Gemini is statistically robust. The three models occupy a relatively narrow performance band, and the choice between Claude and ChatGPT may depend on character-specific strengths documented in Section 3.3.1.

3.3.3. Comparative confusion matrix error analysis

The confusion matrix heatmaps reveal systematic error patterns across all three models (Figure 4). All three show strong diagonal dominance for morphologically distinctive characters such as "rocky cliff," "limestone hill," and "dense palm plantation," confirming that visually salient features are consistently recognised across model architectures.

For infrastructure-related characters, inter-model differences were informative. All three models tended to misclassify "urban setting" as "development view," but this bias was most pronounced in Gemini (9 of 23 cases) compared to ChatGPT (5) and Claude (3), suggesting lower sensitivity to compact urban features.

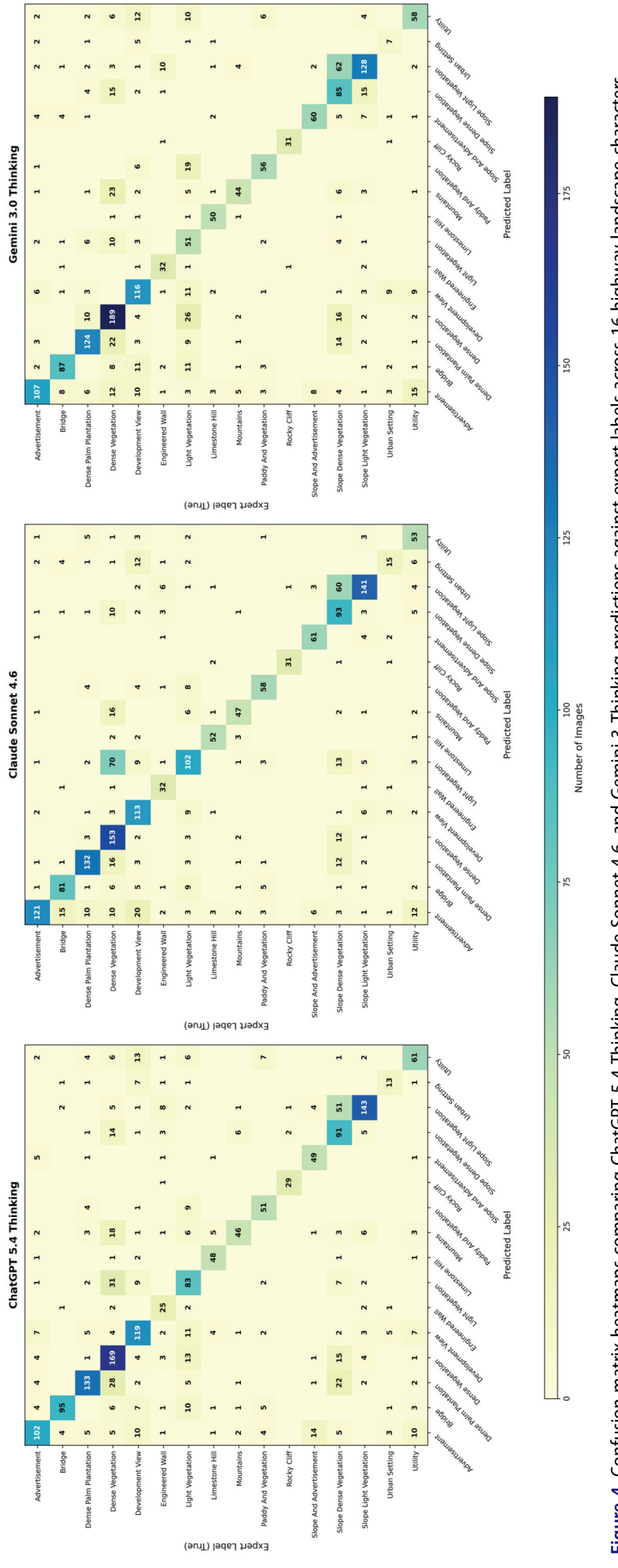


Figure 4. Confusion matrix heatmaps comparing ChatGPT-5.4 Thinking, Claude Sonnet 4.6, and Gemini 3 Thinking predictions against expert labels across 16 highway landscape characters.

For "bridge," ChatGPT achieved the highest recall (95/103), outperforming Gemini (87) and Claude (81). In the "development view," all three performed comparably, though Claude showed notably more misclassifications as "advertisement" (20 cases vs 10 for ChatGPT), while Gemini distributed errors more evenly across "utility" and "bridge."

The most pronounced model differentiation occurred in vegetation-related characters. For "dense vegetation" (289 samples), Gemini achieved the highest recall (189, 65.4%), outperforming ChatGPT (169, 58.5%) and Claude (153, 52.9%). Claude's lower recall was driven by misclassifying "dense vegetation" as "light vegetation" (70 cases – the single largest off-diagonal error across all three matrices, vs 31 for ChatGPT and 10 for Gemini). Conversely, Claude achieved the highest recall for "light vegetation" (102/148 vs 83 and 51), suggesting a systematically lower density threshold for the light – dense boundary. Slope – vegetation differentiation was similarly challenging, with all three models exhibiting substantial misclassification of "slope dense vegetation" as "slope light vegetation" (Claude: 60, Gemini: 62, ChatGPT: 51 of 198 cases), indicating a structurally embedded, model-invariant source of error. "Utility" also proved difficult across all models (Claude: 53/90, ChatGPT: 61/90, Gemini: 58/90), with frequent confusion with "advertisement" reflecting infrastructure partially obscured by roadside greenery.

Taken together, shared errors – slope – vegetation density confusion and the "urban setting"–"development view" overlap – reflect intrinsic taxonomic ambiguities persisting regardless of model choice, while model-specific patterns – Claude's light – dense reclassification bias, ChatGPT's bridge recognition strength, and Gemini's dense vegetation advantage – demonstrate complementary capabilities that justify multi-model triangulation, fulfilling Objective 2.

3.3.4. Inter-model agreement analysis

To complement the per-model analyses above, AI – expert agreement (Cohen's κ) and AI – AI concordance were examined across all three models.

Cohen's κ analysis confirmed Claude's overall superior consistency with expert judgements (κ = 0.68), followed by ChatGPT (0.66) and Gemini (0.64) – all within the substantial agreement range (0.61–0.80) on the Landis and Koch (1977) scale. At the per-class level, Claude achieved the highest κ in 10 of 16 characters, ChatGPT in 4, and Gemini in 3 (Figure 5). The per-class κ distribution largely mirrored the F1 patterns reported in Section 3.3.1: Claude's advantages were concentrated in infrastructure-related and transitional characters (e.g., "slope and advertisement" κ = 0.87 vs. 0.76 and 0.76), ChatGPT showed the strongest expert alignment for "bridge" (0.79) and "urban setting" (0.54), and Gemini led for "rocky cliff" (0.95) and "limestone hill" (0.87), reinforcing the complementary expertise identified earlier.

Pairwise AI – AI concordance analysis revealed moderate inter-model agreement, with Claude – ChatGPT showing the highest concordance (67.9%, κ = 0.65), followed by ChatGPT – Gemini (66.1%, κ = 0.63) and Claude – Gemini (64.3%, κ = 0.61) (Figure 6). Morphologically distinctive characters (e.g., "rocky cliff," "limestone hill") and vegetation-dominated characters (e.g., "slope light vegetation," "dense palm plantation") showed strong diagonal agreement across all three pairs, corresponding precisely to the high-retention tier identified in Section 3.1.

However, systematic inter-model disagreement emerged in semantically ambiguous characters. The most prevalent discordance involved the "light vegetation"–"dense vegetation" boundary, where Claude – Gemini disagreements were particularly frequent (80 cases), compared to Claude – ChatGPT (62) and ChatGPT – Gemini (39) – consistent with Claude's

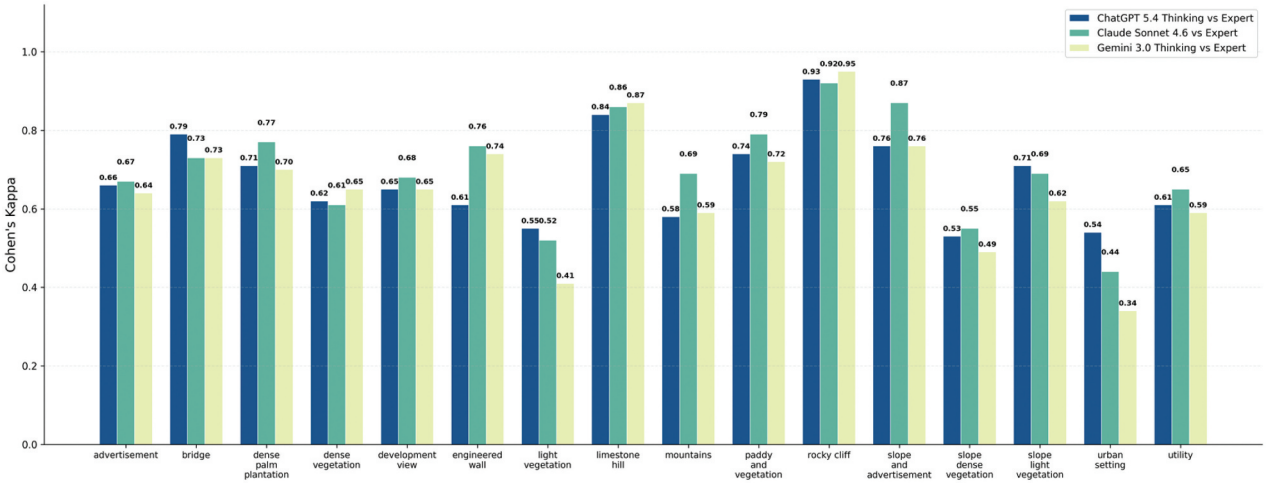


Figure 5. Comparison of Per-Class Cohen's Kappa scores: ChatGPT-5.4 Thinking, Claude Sonnet 4.6, and Gemini 3 Thinking vs expert labels across 16 highway landscape characters.

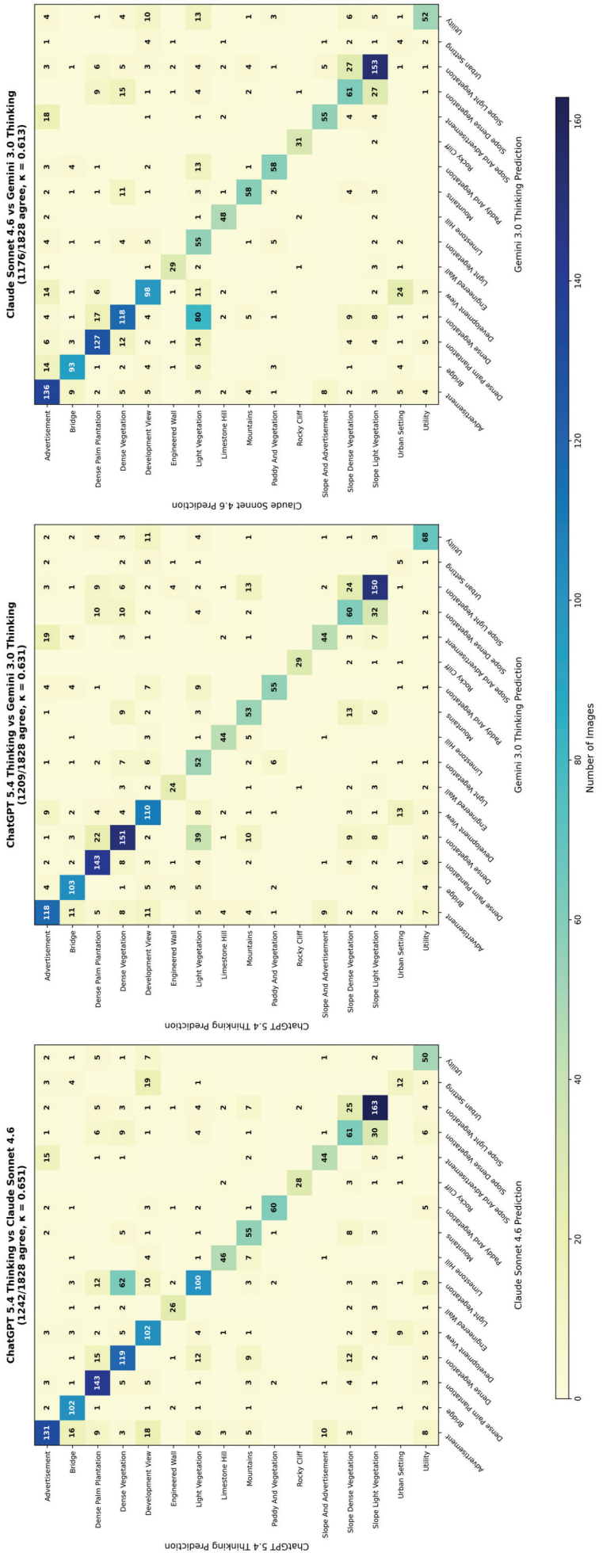


Figure 6. Pairwise AI–AI confusion matrix heatmaps comparing ChatGPT-5.4 Thinking, Claude Sonnet 4.6, and Gemini 3 Thinking predictions across 16 highway landscape characters.

lower density threshold identified in Section 3.3.1. All three model pairs also showed substantial disagreement in slope – vegetation density classification (24–32 cases of "slope light vegetation"–"slope dense vegetation" exchange per pair) and the "urban setting"–"development view" boundary (Claude – Gemini: 24 cases, Claude – ChatGPT: 19, ChatGPT – Gemini: 13), confirming that these represent shared sources of taxonomic ambiguity rather than model-specific weaknesses.

This convergence between AI – expert and AI – AI agreement structures strengthens the framework's capacity to distinguish genuine taxonomic ambiguity from model-specific error, directly fulfilling Objective 3.

4. Discussion

4.1. Methodological contributions in the context of existing Highway landscape research

This study advances highway landscape classification by operationalising an explicitly uncertainty-aware, expert-anchored, AI-assisted validation framework that addresses two methodological gaps: the absence of systematic reliability quantification and the lack of validated AI – expert integration.

As summarised in Section 1.1, existing highway landscape research encompasses context-sensitive road environments, land-use- and landform-based typologies, spatial configuration classifications, and element- or index-based visual quality frameworks. A critical limitation is that class labels are treated as deterministic, with agreement implicit and reliability rarely quantified at the corridor scale. The present framework addresses this gap by introducing a seven-category agreement classification scheme (Categories A – G) that captures the full spectrum of expert – AI concordance across 16 landscape characters – from full four-way agreement (Category A, 45.8%) through various partial agreement configurations to complete disagreement (Category E, 1.2%). Validation thereby shifts from a binary pass/fail check to a graded "reliability surface" that identifies not only what the landscape is, but also how consistently experts and AI models perceive it.

This agreement structure translates into three operational reliability tiers for corridor management: a high-reliability tier (≥90% retention; e.g. "rocky cliff", "limestone hill", "advertisement"), where Categories A and B dominate and images can be used directly with minimal expert checking; an intermediate-reliability tier (70–89%; e.g. "bridge", "dense palm plantation", "mountain"), suitable for preliminary analyses but requiring selective expert review; and a low-reliability tier (<70%; e.g. "urban setting", "utility", "slope dense vegetation", "light vegetation"), where minority consensus and no-consensus cases (Categories C – G) are treated not as outright errors but as systematic indicators of semantic ambiguity triggering targeted expert adjudication. The framework's effectiveness is reflected in an overall retention rate of 74.0% (1,352 of 1,828 images), with the three tiers providing differentiated guidance across the 16-class taxonomy.

Methodologically, the use of zero-shot classification by pre-trained multimodal large language models (MLLMs) such as ChatGPT-5.4 Thinking, Claude Sonnet 4.6, and Gemini 3 Thinking enables effective visual characterisation without task-specific training data, substantially reducing the computational and labour demands of traditional supervised workflows (H. Li et al. 2021). Consistent with recent findings in environmental and perceptual domains (Varga 2025), our results show that foundation models can achieve satisfactory accuracy as "additional raters" within an expert-defined taxonomy (Claude: 70.3%; ChatGPT: 68.8%; Gemini: 67.0%) across diverse highway landscape typologies. For highway agencies managing heterogeneous corridors – from tropical plantations and engineered embankments to karst hillscapes – this multi-model triangulation strategy provides a scalable and transferable protocol that reduces reliance on costly, region-specific model training while preserving expert control over landscape semantics and reliability thresholds.

4.2. Morphological distinctiveness and AI-human alignment

The retention and agreement patterns reported in Sections 3.1–3.2 show that morphological distinctiveness is a primary determinant of classification stability. High-retention characters (>90%) – such as"limestone hill", "rocky cliff", and "slope and advertisement" share three properties: (i) spatially coherent boundaries, (ii) minimal transitional zones, and (iii) distinctive visual signatures. In these classes, all three AI models and the expert frequently converged, with high rates of full agreement and very low rates of disagreement.

These findings align with visual-salience research showing that landscape judgements correlate with edge density, colour saturation, and visual entropy (Celikors and Wells 2022). Characters like "limestone hill" and "rocky cliff" exhibit high-contrast textures, while "dense palm plantation" presents regular vertical patterns-features that create sharp perceptual boundaries. Deep learning research indicates that visual salience is captured by both human vision and convolutional neural networks (Ghariba, Shehata, and McGuire 2019; Kriegeskorte 2015), suggesting that MLLM visual encoders may similarly respond to these characters.

Conversely, low-retention characters (<70%) occupy ecotonal or hybrid positions along continuous gradients. "Urban setting" and "development view" form part of a built-environment spectrum; vegetation density varies along a continuum; and "utility" infrastructure often appears embedded within vegetated or infrastructural matrices. In these cases, minority consensus and no-consensus cases become much more frequent, reflecting semantic ambiguity rather than random error and aligning with disagreement-as-signal perspectives in annotation research (Aroyo and Welty 2013). This echoes the long-standing difficulties in landscape character assessment, as character area boundaries are "rarely precise and generally represent zones of transition" (Tudor 2014).

A particularly informative pattern emerges for "slope dense vegetation", where Category G cases – in which all three AI models converged on the same label yet diverged from the expert – accounted for 15.7% of its samples (31 of 198), the highest share among all characters (Section 3.2). Rather than treating these cases as "AI errors", they can be interpreted as indicative of a systematic boundary misalignment between the expert taxonomy and the models' internal organisation of visual information in slope – vegetation combinations.

4.3. Semantic confusion in AI models and practical implications

Objective 2 compared the performance of ChatGPT-5.4 Thinking, Claude Sonnet 4.6, and Gemini 3 Thinking. As shown in Section 3.3, Claude achieved the highest overall accuracy (70.3%) and macro-averaged F1 score (72.1%), followed by ChatGPT (68.8%, 69.7%) and Gemini (67.0%, 67.6%). Statistical significance testing confirmed that only the Claude – Gemini performance gap was robust (McNemar's $\chi^2 = 6.90$, $p = 0.009$; Wilcoxon $W = 15.0$, $p = 0.004$), while differences between adjacent-ranked models were not statistically significant after Bonferroni correction. Importantly, this performance gap was concentrated in infrastructure-related and transitional environments, whereas all three models performed similarly in morphologically distinctive natural landscapes.

Confusion-matrix analysis (Section 3.3.3) showed that Claude demonstrated the strongest capacity for resolving semantic ambiguities, especially for "urban setting," "development view," and slope – vegetation characters. Claude reduced systematic misclassification of "urban setting" into "development view" relative to both ChatGPT and Gemini, and showed lower confusion between slope – vegetation subcategories than Gemini. However, Claude exhibited a distinctive bias in the light – dense vegetation boundary, misclassifying substantially more "dense vegetation" cases as "light vegetation" (70 cases) than ChatGPT (31) or Gemini (10), suggesting a systematically lower density threshold. Gemini, conversely, achieved the highest recall for "dense vegetation" (65.4%) and morphologically distinctive characters such as "rocky cliff" and "limestone hill," indicating particular responsiveness to high-contrast textures. These patterns contrast with previous benchmarks, where ChatGPT-4o achieved accuracy closest to human ratings, while earlier Claude models exhibited systematic overestimation issues (Varga 2025), suggesting that architectural improvements in Claude Sonnet 4.6 have substantially addressed earlier limitations and translated into more stable performance on infrastructure-heavy and hybrid characters.

The multi-model triangulated validation framework offers three practical roles for highway corridor management. First, for high- and intermediate-reliability classes (e.g. "rocky cliff," "limestone hill," "bridge," "dense vegetation"), the joint AI – expert workflow can pre-classify large volumes of corridor imagery with documented and stratified reliability, reducing manual workload for routine audits, scenic-quality mapping, or ecological-risk screening, consistent with recent studies on AI-assisted environmental monitoring and human – AI collaborative labelling (Yamada et al. 2022; Zurowietz et al. 2018). Second, low-consensus characters – "urban setting," "development view," "slope dense vegetation," and "utility" – are especially relevant in operationally complex segments, because classification uncertainty propagates directly into derived corridor management indicators such as ecological-risk scores (Chang et al. 2023; Zhang, Zheng, et al. 2023) and safety assessments based on street-level imagery (Hanson, Noland, and Brown 2013); low agreement thereby flags locations requiring cautious interpretation or targeted human review. Third, the retained dataset, stratified by agreement type, provides a ready-made training resource: high-agreement images serve as stable prototypes for each landscape character, while minority-consensus and no-consensus cases act as complex examples for learning finer distinctions in ambiguous environments, consistent with recent calls for domain-adapted multimodal models in geographical and environmental applications.

For highway agencies, the three models offer complementary strengths: Claude Sonnet 4.6 should be prioritised for corridors with complex built environments; ChatGPT-5.4 Thinking offers advantages for specific built-environment characters such as bridges and urban settings; while Gemini 3 Thinking performs most strongly on geomorphologically distinctive features. This pattern of distributed expertise supports multi-model triangulation rather than reliance on any single system. Beyond classification, the framework generates actionable

reliability diagnostics that can be integrated into highway landscape grading, maintenance prioritisation, and safety-oriented design.

These complementary strengths can be combined into a single operational framework in three concrete ways. First, character-specific routing and reliability-weighted fusion: the per-class F1 (Section 3.3.1) and Cohen's κ (Section 3.3.4) document which model is most reliable for each character (e.g., Claude for infrastructure-heavy and transitional scenes, Gemini for geomorphologically distinctive features, and ChatGPT for bridges and urban settings); rather than a simple unweighted majority, the three outputs can therefore be combined through a per-class, reliability-weighted vote in which each model's contribution is scaled by its documented reliability for the character it predicts. Second, agreement-based cascading: the A – G scheme itself supplies the control logic, so that full- and majority-consensus images (Categories A – B) are accepted automatically while the remaining images (Categories C – G) are escalated for expert adjudication, concentrating scarce human effort on precisely those scenes where AI and expert judgements fail to converge. Third, complementary error correction: because the models' systematic biases differ – most clearly at the light – dense vegetation boundary, where Claude and Gemini are calibrated to opposite density thresholds – their disagreement can serve as a built-in flag for ambiguous density judgements, triggering either a continuous-index check (Section 4.4) or targeted review. Together, these mechanisms convert the observed distributed expertise into a reproducible multi-model pipeline rather than a post-hoc observation, while keeping experts in control of the final taxonomy.

4.4. Sources of semantic ambiguity and implications for classification design

The low-retention tier reveals three mechanisms of semantic ambiguity with implications for taxonomy design.

First, treating vegetation density as a hard threshold on a continuous gradient yields unstable labels, particularly for slope-related characters. Across both flat and sloping scenes, "light vegetation", "dense vegetation", "slope light vegetation", and "slope dense vegetation" are frequently exchanged (Section 3.3.3, Figure 4), with disagreements clustering around density judgements rather than slope presence. Notably, Claude's tendency to reclassify "dense vegetation" as "light vegetation" (70 cases) while Gemini showed the opposite pattern (only 10 such cases) demonstrates that density thresholds are inconsistently calibrated across AI architectures. These findings suggest that vegetation density would be better represented by continuous or ordinal metrics (e.g., segmentation-derived vegetation fractions, Green View Indices, canopy cover). Such continuous approaches have been successfully applied in street-level greenery assessment (X. Li et al. 2015), and streetscape greenery has been shown to exhibit non-linear, threshold-dependent associations with walking propensity among older adults (Yang et al. 2021), implying that continuous metrics may reveal behavioural relationships that categorical approaches would obscure.

Second, overlapping categorical domains characterise "urban setting" and "development view", both of which represent human-dominated landscapes but emphasise different aspects. All three models showed a tendency to conflate these characters, with Gemini's bias being particularly pronounced (9 of 23 urban setting cases misclassified as development view). A hierarchical taxonomy – nesting both under a broader "developed landscape" class with additional attributes – would better reflect their conceptual relationship, consistent with hierarchical landscape-character protocols in LCA practice (Tudor 2014).

Third, context-dependent visual salience affects "utility" infrastructure, which may be visually dominant in open settings but secondary in dense vegetation. In our results, "utility" is misclassified across a wide range of classes, reflecting its typical role as a secondary landscape element. Treating utility as a feature attribute attached to a primary character (e.g., "dense vegetation with utility") would be more robust and consistent with "landscape unit + feature" schemes in LCA methodologies.

These three mechanisms provide a conceptual bridge between empirical results and practical scheme design: high-retention characters can be prioritised as stable indicators, while low-retention characters should be re-defined, embedded within hierarchical systems, or supplemented with continuous metrics. The multi-model triangulated framework thereby reveals where class boundaries are intrinsically fragile and provides concrete guidance for restructuring future highway landscape classification systems.

4.5. Limitations and future studies

Several limitations constrain the generalisability of the present findings. First, the empirical analysis is restricted to a single 418-km corridor along Malaysia's North-South Expressway. While the empirical results are corridor-specific and may require recalibration for different climatic regions, cultural contexts, or infrastructure typologies (Bürgi et al. 2017), the multi-model triangulated framework itself – built on a three-dimensional classification basis (landform, land use and land cover) with

generic character labels – is conceptually transferable. The reliability tiers reported here should be interpreted as a first empirical benchmark; independent applications to highways in other contexts are required to test the framework's boundary conditions.

Second, the consensus mechanism is built on one human expert and three proprietary ML models. Although three models enable genuine triangulation and richer agreement diagnostics (the seven-category A – G scheme), the framework does not yet realise the full variance-reduction benefits of larger ensembles (e.g. five or more learners). Additional models – including open-source vision – language systems such as LLaVA and Grok (S. Yin et al. 2024) – may exhibit different strengths and biases. Furthermore, we did not benchmark the zero-shot MLLMs against a dedicated supervised classifier trained on the same expert labels, which might achieve higher accuracy for a purely predictive objective. However, our primary goal was to demonstrate an expert-anchored, uncertainty-aware validation framework rather than to identify the single most accurate architecture. Future work should therefore (i) expand triangulation to five or more models, and (ii) complement the multi-model design with supervised baselines, quantifying trade-offs between task-specific optimisation and the scalability offered by general-purpose MLLMs (Mewada et al. 2025; Wortsman et al. 2022).

Third, all classifications were conducted on single, static frames, despite the perception of the highway landscape being sequential. Incorporating short image sequences, GPS coordinates, elevation data, and land-use layers could enhance the disambiguation of transitional classes. Additionally, implementing explainable AI tools (e.g., attention maps, Grad-CAM) would reveal which visual cues drive AI decisions, enabling a systematic audit of whether AI systems attend to expert-relevant features (Qi et al. 2025).

Fourth, while expert annotations anchored the taxonomy, inter-expert variability was not explicitly modelled. This limits the ability to distinguish AI – human disagreement from human – human disagreement. We mitigated this by grounding the taxonomy in a transparent three-dimensional framework and interpreting disagreement within the multi-model expert – AI structure. Notably, Category G – where all three AI models converge on an alternative label – highlights cases where expert labels may warrant reconsideration, transforming the validation from unidirectional benchmarking into a reflexive calibration process. Since disagreement among human raters is an inherent feature of landscape evaluation (Clay and Smidt 2004; Palmer 2000), future work should include multi-expert panels with formal inter-rater reliability analysis, positioning model performance relative to realistic human baselines rather than idealised single-expert benchmarks.

By addressing these limitations, future research can expand the present framework into adaptive, context-sensitive tools for highway landscape assessment, supporting visual-quality monitoring, safety-oriented design, and ecosystem-aware corridor planning at the network scale.

4.6. Scaling from a single corridor to highway networks and urban streets

Although the present analysis is confined to a single corridor, the framework is explicitly designed for extension to highway networks. Because the three-dimensional basis (landform, land use, and land cover) and the generic character labels are not tied to any single route, the taxonomy and the A – G agreement scheme can be transferred directly to additional corridors, with experts re-anchoring only the region-specific characters that arise under different climatic, vegetative, or infrastructural conditions. Systematic sampling at the network scale need not rely on new field campaigns: street-level imagery services such as Google Street View (and comparable platforms) can be queried programmatically to extract georeferenced views at fixed intervals across an entire network. Within such a deployment, the reliability tiers established here (Section 4.1) function as a triage mechanism – high-reliability characters are auto-processed with minimal oversight, whereas low-reliability segments are automatically flagged for targeted expert review – allowing agencies to allocate limited verification effort efficiently across thousands of kilometres.

Extending the framework to dense urban street networks, as suggested, is a more demanding but particularly promising direction that would require three adaptations. First, the higher spatial frequency of change along city streets calls for denser sampling than the 250 m interval used here. Second, the driver-centric taxonomy would need re-anchoring around pedestrian- and cyclist-oriented characters (e.g., street-canyon enclosure, active frontages, street furniture) and would ideally accommodate multiple user viewpoints rather than a single driver perspective. Third, the density-gradient ambiguities identified in Section 4.4 become even more acute in dense streetscapes, where the continuous or ordinal greenery indices proposed there – such as the Green View Index – would integrate naturally with the agreement scheme and are especially pertinent to pedestrian-oriented visual quality. Crucially, the triangulated expert – AI agreement structure is itself domain-agnostic: it can wrap any street-level taxonomy to deliver the same graded reliability stratification,

supporting network-scale mapping of urban visual quality, walkability, and safety.

5. Conclusions

This study developed and validated a multi-model expert – AI triangulated validation framework for highway landscape classification. Across 1,828 images from Malaysia's North-South Expressway, Claude Sonnet 4.6 achieved the highest overall accuracy (70.3%), followed by ChatGPT-5.4 Thinking (68.8%) and Gemini 3 Thinking (67.0%), with only the Claude – Gemini performance gap reaching statistical significance ($p < 0.05$). The agreement-weighted multi-model approach retained 74.0% of samples (1,352 images) with high confidence. Classification reliability varied systematically: morphologically distinct landscapes achieved over 90% retention, whereas semantically ambiguous characters exhibited persistent disagreement, revealing fragile taxonomic boundaries and clarifying the boundary conditions of zero-shot multimodal AI in landscape assessment.

The framework advances existing methodology in three ways. Methodologically, it reframes validation from a deterministic, single-label assignment to an uncertainty-aware, graded-reliability surface that treats MLLMs as additional raters rather than replacements for experts. Empirically, it identifies specific landscape characters where AI classification is trustworthy versus where expert adjudication remains essential, with the three-model comparison revealing complementary strengths – Claude in infrastructure-heavy environments, ChatGPT in specific built-environment characters, and Gemini in geomorphologically distinctive features – demonstrating that multi-model triangulation captures diagnostic information that no single model provides. Practically, it generates a quality-controlled dataset of 1,352 images stratified by agreement type, supporting corridor management, visual-quality monitoring, and future region-specific model fine-tuning. By embedding AI within an expert-anchored workflow, this approach provides a scalable, reproducible protocol for landscape assessment that balances automation with interpretive expertise.

Future research should expand the framework to more diverse geographic and climatic contexts, incorporate spatial and temporal information, include additional AI models and open-source vision – language systems, and explore domain-specific fine-tuning. These extensions would strengthen the framework's generalisability beyond single-corridor applications. Overall, this research establishes a practical methodology for AI – human collaboration in landscape assessment, providing a scalable approach to generating validated datasets and supporting evidence-based decision-making in transportation infrastructure planning.

Author contributions

CRediT: **Hangyu Gao:** Conceptualization, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing; **Richard Smardon:** Methodology, Software, Validation, Writing – review & editing; **Shamsul Abu Bakar:** Conceptualization, Formal analysis, Methodology; **Suhardi Maulan:** Methodology, Software, Validation; **Jiani Yang:** Conceptualization, Writing – review & editing; **Riyadh Mundher:** Visualization; **Yijiang Zhou:** Data curation.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The author(s) reported there is no funding associated with the work featured in this article.

Data availability statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Anthropic. 2026. "Introducing Claude Sonnet 4.6." Feb 16. Accessed March 14, 2026, from https://www.anthropic.com/news/claude-sonnet-4-6.

Aroyo, L., and C. Welty. 2013. "Crowd Truth: Harnessing Disagreement in Crowdsourcing a Relation Extraction Gold Standard." OPAL (Open@LaTrobe) (La Trobe University). https://doi.org/10.6084/m9.figshare.679997.v1.

Atik, M., R. C. Işıklı, V. Ortaçeşme, and E. Yıldırım. 2017. "Exploring a Combination of Objective and Subjective Assessment in Landscape Classification: Side Case from Turkey." *Applied Geography* 83:130–140. https://doi.org/10.1016/j.apgeog.2017.04.004.

Bai, Z., R. Ji, and J. Qi. 2024. "Deciphering motorists' Perceptions of Scenic Road Visual Landscapes: Integrating Binocular Simulation and Image Segmentation." *Land* 13 (9): 1381. https://doi.org/10.3390/land13091381.

Bastian, O. 2008. "Landscape Classification Between Fact and Fiction." Paper presented at the conference Warzaw University and Polish Association of Landscape Ecology, Warzawa. Polska Asocjacja Ekologii Krajobrazu.

Brabyn, L. 2009. "Classifying Landscape Character." *Landscape Research* 34 (3): 299–321. https://doi.org/10.1080/01426390802371202.

Brown, G. 2003. "A Method for Assessing Highway Qualities to Integrate Values in Highway Planning." *Journal of*

Transport Geography 11 (4): 271–283. https://doi.org/10.1016/s0966-6923(03)00004-8.

Bürgi, M., C. Bieling, K. Von Hackwitz, T. Kizos, J. Lieskovský, M. G. Martín, S. McCarthy, et al. 2017. "Processes and Driving Forces in Changing Cultural Landscapes Across Europe." *Landscape Ecology* 32 (11): 2097–2112. https://doi.org/10.1007/s10980-017-0513-z.

Celikors, E., and N. M. Wells. 2022. "Are Low-Level Visual Features of Scenes Associated with Perceived Restorative Qualities?" *Journal of Environmental Psychology* 81:101800. https://doi.org/10.1016/j.jenvp.2022.101800.

Chang, S., Z. Dai, X. Wang, Z. Zhu, and Y. Feng. 2023. "Landscape Pattern Identification and Ecological Risk Assessment Employing Land Use Dynamics on the Loess Plateau." *Agronomy* 13 (9): 2247. https://doi.org/10.3390/agronomy13092247.

Clay, G. R., and R. K. Smidt. 2004. "Assessing the Validity and Reliability of Descriptor Variables Used in Scenic Highway Analysis." *Landscape and Urban Planning* 66 (4): 239–255. https://doi.org/10.1016/S0169-2046(03)00114-2.

Gao, H., S. A. Bakar, S. Maulan, M. J. M. Yusof, R. Mundher, and B. Chen. 2024. "How Highway Landscape Visual Qualities are Being Studied: A Systematic Literature Review." *Land* 13 (4): 431. https://doi.org/10.3390/land13040431.

Gao, H., S. A. Bakar, S. Maulan, M. J. M. Yusof, R. Mundher, and K. Zakariya. 2023. "Identifying Visual Quality of Rural Road Landscape Character by Using Public Preference and Heatmap Analysis in Sabak Bernam, Malaysia." *Land* 12 (7): 1440. https://doi.org/10.3390/land12071440.

Gao, H., S. A. Bakar, M. Suhardi, Y. Guo, M. J. M. Yusof, R. Mundher, Y. Zhuo, and J. Qi. 2025. "Constructing a Conceptual Framework: Interpreting Visual Preference and Visual Pollution Factors Among Viewers in Highway Landscapes." *Transportation Research Interdisciplinary Perspectives* 31:101399. https://doi.org/10.1016/j.trip.2025.101399.

Geiger, R. S., D. Cope, J. Ip, M. Lotosh, A. Shah, J. Weng, and R. Tang. 2021. "Garbage In, Garbage Out" Revisited: What Do Machine Learning Application Papers Report About Human-Labeled Training Data?" *Quantitative Science Studies* 2 (3): 795–827. https://doi.org/10.1162/qss_a_00144.

Ghariba, B., M. S. Shehata, and P. McGuire. 2019. "Visual Saliency Prediction Based on Deep Learning." *Information* 10 (8): 257. https://doi.org/10.3390/info10080257.

Google. 2025. "A new era of intelligence with Gemini 3." November 18. Accessed March 14, 2026, from https://blog.google/products-and-platforms/products/gemini/gemini-3/.

Grazuleviciute-Vileniske, I., and I. Matijosaitiene. 2010. "Cultural Heritage of Roads and Road Landscapes: Classification and Insights on Valuation." *Landscape Research* 35 (4): 391–413. https://doi.org/10.1080/01426397.2010.486856.

Guo, F., M. Li, Y. Chen, J. Xiong, and J. Lee. 2019. "Effects of Highway Landscapes on drivers' Eye Movement Behavior and Emergency Reaction Time: A Driving Simulator Study." *Journal of Advanced Transportation* 2019:1–9. https://doi.org/10.1155/2019/9897831.

Hallo, J. C., and R. E. Manning. 2009. "Transportation and Recreation: A Case Study of Visitors Driving for Pleasure at Acadia National Park." *Journal of Transport Geography* 17 (6): 491–499. https://doi.org/10.1016/j.jtrangeo.2008.10.001.

Hanson, C. S., R. B. Noland, and C. Brown. 2013. "The Severity of Pedestrian Crashes: An Analysis Using Google Street View Imagery." *Journal of Transport Geography* 33:42–53. https://doi.org/10.1016/j.jtrangeo.2013.09.002.

Hu, S., G. He, and H. Xu. 2023. "Research on the Segmentation Method of a Highway Landscape Corridor Zone." *Frontiers in Environmental Science* 11:1212264. https://doi.org/10.3389/fenvs.2023.1212264.

Huijuan, L., H. Teng, M. Xinyu, and C. Baodong. 2025. "The Impact of Transportation Infrastructure on the Regional Economic Integration in China: A CGE Analysis." *International Review of Economics and Finance* 99:104045. https://doi.org/10.1016/j.iref.2025.104045.

Jaal, Z., J. Abdullah, and H. Ismail. 2013. "Malaysian North South Expressway Landscape Character: Analysis of users' Preference of Highway Landscape Elements." *WIT Transactions on Ecology and the Environment* 1:365–376. https://doi.org/10.2495/SC130311Qi.

James, P., and J. W. Gittins. 2007. "Local Landscape Character Assessment: An Evaluation of Community-Led Schemes in Cheshire." *Landscape Research* 32 (4): 423–442. https://doi.org/10.1080/01426390701449794.

Jellema, A., D. Stobbelaar, J. C. Groot, and W. A. Rossing. 2009. "Landscape Character Assessment Using Region Growing Techniques in Geographical Information Systems." *Journal of Environmental Management* 90:S161–S174. https://doi.org/10.1016/j.jenvman.2008.11.031.

Jiang, B., J. He, J. Chen, and L. Larsen. 2021. "Moderate is Optimal: A Simulated Driving Experiment Reveals Freeway Landscape Matters for Driving Performance." *Urban Forestry and Urban Greening* 58:126976. https://doi.org/10.1016/j.ufug.2021.126976.

Jiang, P., H. Wu, Y. Zhao, D. Zhao, G. Zhou, and C. Xin. 2024. "SEEK+: Securing Vehicle GPS via a Sequential Dashcam-Based Vehicle Localisation Framework." *Pervasive and Mobile Computing* 100:101916. https://doi.org/10.1016/j.pmcj.2024.101916.

Kaul, V., S. Enslin, and S. A. Gross. 2020. "History of Artificial Intelligence in Medicine." *Gastrointestinal Endoscopy* 92 (4): 807–812. https://doi.org/10.1016/j.gie.2020.06.040.

Kazmierczak, R., E. Berthier, G. Frehse, and G. Franchi. 2025. "Explainability and Vision Foundation Models: A Survey." *Information Fusion* 122:103184. https://doi.org/10.1016/j.inffus.2025.103184.

Kelly, C. M., J. S. Wilson, E. A. Baker, D. K. Miller, and M. Schootman. 2012. "Using Google Street View to Audit the Built Environment: Inter-Rater Reliability Results." *Annals of Behavioral Medicine* 45 (S1): 108–112. https://doi.org/10.1007/s12160-012-9418-9.

Kent, R. L. 1993. "Determining Scenic Quality Along Highways: A Cognitive Approach." *Landscape and Urban Planning* 27 (1): 29–45. https://doi.org/10.1016/0169-2046(93)90026-A.

Koyun, M., and I. Taskent. 2025. "Evaluation of Advanced Artificial Intelligence algorithms' Diagnostic Efficacy in Acute Ischemic Stroke: A Comparative Analysis of ChatGPT-4o and Claude 3.5 Sonnet Models." *Journal of Clinical Medicine* 14 (2): 571. https://doi.org/10.3390/jcm14020571.

Kriegeskorte, N. 2015. "Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing." *Annual Review of Vision Science* 1 (1): 417–446. https://doi.org/10.1146/annurev-vision-082114-035447.

Landis, J. R., and G. G. Koch. 1977. "An Application of Hierarchical Kappa-Type Statistics in the Assessment of Majority Agreement Among Multiple Observers." *Biometrics* 33 (2): 363. https://doi.org/10.2307/2529786.

Larue, G. S., A. Rakotonirainy, and A. N. Pettitt. 2011. "Driving Performance Impairments Due to Hypovigilance on Monotonous Roads." *Accident Analysis and Prevention* 43 (6): 2037–2046. https://doi.org/10.1016/j.aap.2011.05.023.

Li, H., C. Zhang, S. Zhang, X. Tan, X. Ding, and P. M. Atkinson. 2021. "Iterative Deep Learning (IDL) for Agricultural Landscape Classification Using Fine Spatial Resolution Remotely Sensed Imagery." *International Journal of Applied Earth Observation and Geoinformation* 102:102437. https://doi.org/10.1016/j.jag.2021.102437.

Li, K., G. Vosselman, and M. Y. Yang. 2025. "Multimodal Rationales for Explainable Visual Question Answering." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*, Nashville, TN, USA, 191–201. Los Alamitos, CA: IEEE Computer Society.

Li, X., C. Zhang, W. Li, R. Ricard, Q. Meng, and W. Zhang. 2015. "Assessing Street-Level Urban Greenery Using Google Street View and a Modified Green View Index." *Urban Forestry and Urban Greening* 14 (3): 675–685. https://doi.org/10.1016/j.ufug.2015.06.006.

Martín, B., R. Arce, I. Otero, and M. Loro. 2018. "Visual Landscape Quality as Viewed from Motorways in Spain." *Sustainability* 10 (8): 2592. https://doi.org/10.3390/su10082592.

Merry, K., P. Bettinger, J. Siry, and J. Bowker. 2019. "Preferences of Motorcyclists to Views of Managed, Rural Southern United States Landscapes." *Journal of Outdoor Recreation and Tourism* 29:100259. https://doi.org/10.1016/j.jort.2019.100259.

Mewada, D., E. M. Grua, C. Eising, P. Denny, P. Van De Ven, and A. Scanlan. 2025. "Zero-Shot Learning for Sustainable Municipal Waste Classification." *Recycling* 10 (4): 144. https://doi.org/10.3390/recycling10040144.

Northcutt, C., L. Jiang, and I. Chuang. 2021. "Confident Learning: Estimating Uncertainty in Dataset Labels." *The Journal of Artificial Intelligence Research* 70:1373–1411. https://doi.org/10.1613/jair.1.12125.

Novack, Z., J. McAuley, Z. C. Lipton, and S. Garg. 2023. "Chils: Zero-Shot Image Classification with Hierarchical Label Sets." In *Proceedings of the 40th International Conference on Machine Learning, Proceedings of Machine Learning Research 202*, edited by A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, 26342–26362. PMLR.

OpenAI. 2026. "Introducing GPT-5.4." March 5. Accessed March 14, 2026, from https://openai.com/zh-Hans-CN/index/introducing-gpt-5-4/.

Palmer, J. F. 2000. "Reliability of Rating Visible Landscape Qualities." *Landscape Journal* 19 (1–2): 166–178. https://doi.org/10.3368/lj.19.1-2.166.

Qi, J., W. Li, Z. Bai, H. Gao, X. Tang, and Y. Zhou. 2025. "Estimating Aesthetic Services of Road Landscapes Through Predicting people's Attention: A Computer Vision Approach." *Journal of Environmental Management* 376:124584. https://doi.org/10.1016/j.jenvman.2025.124584.

Qin, X., M. Fang, D. Yang, and V. W. Wangari. 2023. "Quantitative Evaluation of Attraction Intensity of Highway Landscape Visual Elements Based on Dynamic Perception." *Environmental Impact Assessment Review* 100:107081. https://doi.org/10.1016/j.eiar.2023.107081.

Qin, X., Y. Liu, D. Yang, D. Pan, F. Meng, Y. Cao, and V. W. Wangari. 2024. "Quantitative Characterization of Highway Landscape Space Visual Perception Based on Deep Learning." *IEEE Transactions on Intelligent Transportation Systems* 25 (12): 21157–21171. https://doi.org/10.1109/TITS.2024.3454482.

Qin, X., D. Yang, and V. W. Wangari. 2024. "Quantitative Characterization and Evaluation of Highway Greening Landscape Spatial Quality Based on Deep Learning." *Environmental Impact Assessment Review* 107:107559. https://doi.org/10.1016/j.eiar.2024.107559.

Radford, A., J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, and I. Sutskever. 2021. "Learning Transferable Visual Models from Natural Language Supervision." In *Proceedings of the 38th International Conference on Machine Learning, Proceedings of Machine Learning Research 139*, edited by M. Meila and T. Zhang, 8748–8763. PMLR.

Simensen, T., R. Halvorsen, and L. Erikstad. 2018. "Methods for Landscape Characterisation and Mapping: A Systematic Review." *Land Use Policy* 75:557–569. https://doi.org/10.1016/j.landusepol.2018.04.022.

Stamatiadis, N., A. Kirk, and L. Wright. 2023. "Context Classification: A New Approach for Improving Highway Design." *Transportation Research Procedia* 69:45–52. https://doi.org/10.1016/j.trpro.2023.02.143.

Tan, C., Q. Cao, Y. Li, J. Zhang, X. Yang, H. Zhao, Z. Wu, et al. 2023. "On the Promises and Challenges of Multimodal Foundation Models for Geographical, Environmental, Agricultural, and Urban Planning Applications." arXiv. https://doi.org/10.48550/arxiv.2312.17016.

Tang, I., and T. P. Breckon. 2011. "Automatic Road Environment Classification." *IEEE Transactions on Intelligent Transportation Systems* 12 (2): 476–484. https://doi.org/10.1109/TITS.2010.2095499.

Terkenli, T., A. Gkoltsiou, and D. Kavroudakis. 2021. "The Interplay of Objectivity and Subjectivity in Landscape Character Assessment: Qualitative and Quantitative Approaches and Challenges." *Land* 10 (1): 53. https://doi.org/10.3390/land10010053.

Thiffault, P., and J. Bergeron. 2003. "Monotony of Road Environment and Driver Fatigue: A Simulator Study." *Accident Analysis and Prevention* 35 (3): 381–391. https://doi.org/10.1016/S0001-4575(02)00014-3.

Tudor, C. 2014. "An Approach to Landscape Character Assessment." *Natural England* 65:101716.

Varga, D. 2025. "Comparative Evaluation of Multimodal Large Language Models for No-Reference Image Quality Assessment with Authentic Distortions: A Study of OpenAI and Claude.AI Models." *Big Data and Cognitive Computing* 9 (5): 132. https://doi.org/10.3390/bdcc9050132.

Wortsman, M., G. Ilharco, J. W. Kim, M. Li, S. Kornblith, R. Roelofs, R. G. Lopes, et al. 2022. "Robust Fine-Tuning of Zero-Shot Models." In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, edited by R. Chellappa, 7949–7961. Los Alamitos, CA: IEEE Computer Society. https://doi.org/10.1109/CVPR52688.2022.00780.

Yamada, T., M. Massot-Campos, A. Prugel-Bennett, O. Pizarro, S. B. Williams, and B. Thornton. 2022. "Guiding Labelling Effort for Efficient Learning with Georeferenced Images." *IEEE Transactions on Pattern Analysis & Machine Intelligence* 45 (1): 593–607. https://doi.org/10.1109/TPAMI.2021.3140060.

Yang, L., Y. Ao, J. Ke, Y. Lu, and Y. Liang. 2021. "To Walk or Not to Walk? Examining Non-Linear Effects of Streetscape Greenery on Walking Propensity of Older Adults." *Journal of Transport Geography* 94:103099. https://doi.org/10.1016/j.jtrangeo.2021.103099.

Yin, L., Q. Cheng, Z. Wang, and Z. Shao. 2015. "'Big data' For Pedestrian Volume: Exploring the Use of Google Street View Images for Pedestrian Counts." *Applied Geography* 63:337–345. https://doi.org/10.1016/j.apgeog.2015.07.010.

Yin, S., C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen. 2024. "A Survey on Multimodal Large Language Models." *National Science Review* 11 (12): nwae403. https://doi.org/10.1093/nsr/nwae403.

Zainon, M., M. Z. Shah, M. A. Chiroma, and M. A. Kafi. 2018. "Monotonous Driving Environment Along Highway and Driver Behaviour in Malaysia: A Review." *Journal of the Society of Automotive Engineers Malaysia* 2 (1): 60–74. https://doi.org/10.56381/jsaem.v2i1.78.

Zhai, X., X. Wang, B. Mustafa, A. Steiner, D. Keysers, A. Kolesnikov, and L. Beyer. 2022. "LiT: Zero-Shot Transfer with Locked-Image Text Tuning." In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, edited by R. Chellappa, 18102–18112. Los Alamitos, CA: IEEE Computer Society. https://doi.org/10.1109/CVPR52688.2022.01759.

Zhang, J., R. Hu, X. Cheng, V. Christos, S. P. Philbin, R. Zhao, and X. Zhao. 2023. "Assessing the Landscape Ecological Risk of Road Construction: The Case of the Phnom Penh-Sihanoukville Expressway in Cambodia." *Ecological Indicators* 154:110582. https://doi.org/10.1016/j.ecolind.2023.110582.

Zhang, J., S. Zheng, Y. Sun, and H. Yue. 2023. "Landscape Ecological Risk Assessment of an Ecological Area in the Kubuqi Desert Based on Landsat Remote Sensing Data." *PLOS ONE* 18 (11): e0294584. https://doi.org/10.1371/journal.pone.0294584.

Zurowietz, M., D. Langenkämper, B. Hosking, H. A. Ruhl, and T. W. Nattkemper. 2018. "MAIA—A Machine Learning Assisted Image Annotation Method for Environmental Monitoring and Exploration." *PLOS ONE* 13 (11): e0207498. https://doi.org/10.1371/journal.pone.0207498.