land

MDPI

*Article*

# Highway Landscape Preference Along Malaysia's North–South Expressway: A Comparison of Multimodal Large Language Models and Human Judgments

**Hangyu Gao [1], Richard Smardon [2,*], Shamsul Abu Bakar [1], Suhardi Maulan [1] and Jiani Yang [3]**

[1] Department of Landscape Architecture, Faculty of Design and Architecture, Universiti Putra Malaysia, Serdang 43400, Malaysia; gs58413@student.upm.edu.my (H.G.); shamsul_ab@upm.edu.my (S.A.B.); suhardi@upm.edu.my (S.M.)
[2] Department of Environmental Studies, SUNY College of Environmental Science and Forestry, 1 Forestry Drive, Syracuse, NY 13210, USA
[3] Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, CA 91125, USA; yjn@caltech.edu
* Correspondence: rsmardon@esf.edu

**Abstract**

Visual landscape assessment informs highway corridor planning decisions, yet conventional surveys scale poorly with corridor length. Multimodal large language models (MLLMs) offer a scalable alternative, but their alignment with road-user preferences remains poorly understood. This study aimed to quantify the extent to which MLLM-derived visual preference rankings align with those of road users. The comparison used 80 images across 16 landscape character groups along 418 km of Malaysia's North–South Expressway. Five MLLMs (ChatGPT, Claude, Gemini, Kimi, and Qwen) were queried under three prompt formulations using complete pairwise comparison with AB/BA reversal, yielding 94,800 judgements. Bradley–Terry rankings were then compared against rating-scale responses from 400 road users. The five models converged strongly (Kendall's W = 0.926), whereas baseline AI–human agreement was moderate (image-level ρ = 0.622; group-level ρ = 0.697). Divergences are concentrated in two opposing categories. Paddy landscapes, ranked first by humans, fell to thirteenth in the AI ranking, whereas advertisement-dominated scenes were overvalued. Excluding the paddy group raised correlations to 0.772 and 0.911. A theory-directed prompt achieved comparable gains (ρ = 0.775 and 0.929) and restored paddy to third rank. A hybrid AI-screening, human-targeted protocol is proposed for corridor-scale visual planning.

**Keywords:** multimodal large language model; highway landscape; pairwise comparison; Bradley–Terry model; prompt engineering; landscape character assessment; road-user perception



Academic Editor: Mark Altaweel

Received: 5 June 2026
Revised: 7 July 2026
Accepted: 8 July 2026
Published: 9 July 2026



## 1. Introduction

Highway corridor planning increasingly requires systematic attention to the visual environment that road users encounter at speed. As primary arteries of regional connectivity, these corridors expose vast numbers of users to extended sequences of visual environments [1]. The roadside landscape thus serves as a consequential component of the public realm, where users encounter diverse visual elements, including natural terrain, vegetation, engineered structures, and cultural landmarks [2].

Appreciating the aesthetic qualities of highway landscapes, including their contribution to regional identity and cultural heritage [3,4], is complicated by the unique perceptual conditions of highway travel. Unlike urban or parkland settings, where users move slowly and can freely redirect their gaze, highway travel imposes a constrained forward-facing perspective at high speed, affording limited temporal windows for visual fixation and cognitive processing of roadside elements [5,6].

Moreover, highway corridors traverse heterogeneous land-use mosaics that may transition rapidly from agricultural landscapes to suburban developments, industrial zones, and urban cores within a single journey [7]. This spatial heterogeneity, combined with the temporal compression inherent in high-speed travel, creates a dynamic perceptual environment that cannot be adequately captured by traditional static landscape assessment methodologies developed for parks, urban squares, or recreational areas.

Contemporary urbanisation has fundamentally altered the visual character of highway corridors [8] as suburban development progressively encroaches on formerly rural and natural landscapes. Numerous infrastructure projects have proceeded with inadequate attention to landscape ecological consequences, resulting in land-use conversion, vegetation loss, and increased landscape fragmentation [9]. Understanding how highway users perceive these evolving environments is therefore essential for corridor management strategies that balance infrastructure needs with landscape quality.

Research on landscape perception has long recognised that such evaluations arise from the interaction between biophysical landscape properties and the perceptual–cognitive processes of human observers, a relationship that has been studied through expert, psychophysical, cognitive, and experiential paradigms [10]. Within the psychophysical tradition, in particular, photo-based survey methods have become a standard tool in which representative images are presented to respondents who rate their visual appeal on ordinal scales, and these ratings are then linked to measurable landscape features [11]. Such approaches have proven effective for establishing preference hierarchies among landscape types and for identifying structural attributes—such as land-type diversity and the presence of water—that predict visual preference [12].

Within Southeast Asia, and Malaysia in particular, this photo-based tradition has documented distinctive preferences for tropical expressway scenery. Early reviews of Malaysian highway landscapes identified the Kedah paddy plains and the Kinta Valley limestone hills as emblematic scenic assets of the national expressway network [13], and user surveys along the North–South Expressway itself confirmed strong preferences for natural and cultivated roadside elements [14]. The cultural weight of these agricultural scenes is well established: across Southeast Asia, traditional rice landscapes carry deep associations with cultural identity and landscape aesthetics [15].

However, these survey-based methods remain resource-intensive, time-consuming, and difficult to scale across broader spatial contexts [16]. These logistical demands escalate sharply with corridor length, limiting the spatial coverage that human surveys alone can achieve along extended highway networks. The recent emergence of multimodal large language models (MLLMs), capable of processing both images and text, has raised the prospect that such models could serve as complementary tools for simulating human landscape preference evaluations at scale. However, their alignment with the nuanced judgements of different stakeholder groups requires further validation [17].

Recent advances in MLLMs have opened new possibilities for automated visual assessment. Unlike earlier computational approaches that relied on curated aesthetic datasets and task-specific pipelines [18], contemporary MLLMs integrate visual and linguistic processing within unified architectures trained on web-scale corpora, enabling flexible vision-language reasoning across diverse tasks [19]. Emerging applications have shown that MLLMs can

score landscapes against predefined criteria with moderate positive correlations to human ratings, though systematic geographic biases persist [20]. Assessments of natural landscape preferences similarly indicate significant AI–human alignment on dimensions such as complexity, coherence, mystery, and overall preference, but not on legibility [17]. However, findings remain inconsistent: Stakeholder-informed prompt engineering can improve model–human correlation by 38–85% [16], and baseline MLLM predictions often diverge markedly from human rankings [21]. Moreover, broader evaluations have revealed a persistent gap between perceptual accuracy and genuine cognitive understanding in MLLMs, with models failing a substantial proportion of reasoning tasks even when visual recognition is correct [22]. This variability underscores the need to move beyond global agreement metrics and identify the specific landscape conditions under which AI–human convergence holds or breaks down.

Several gaps limit the current understanding of MLLM-based landscape preference assessment. First, studies of landscape preference have relied exclusively on absolute rating formats, in which models assign numerical scores to individual images [16,17,21]. This approach is susceptible to calibration instability across judge models [23] and cannot diagnose position bias—the tendency to favour whichever option appears first or second in a comparison—a limitation extensively documented in text-based LLM-as-a-judge tasks [23,24]. Pairwise comparison has recently been introduced in adjacent urban-perception research, but existing implementations either employed a single model [25,26] or, where multiple models were compared, sampled an incomplete subset of pairs for safety perception without positional counterbalancing [27]. No visual landscape study has yet to implement a complete pairwise design with systematic positional reversal, leaving the position bias unquantified in this domain. Second, prior work has focused on single landscape or environmental contexts, such as agricultural heritage sites [16], natural scenes from standardised datasets [17], urban green and blue spaces [21], or street-level urban environments [20,25,27,28], without systematically evaluating AI–human agreement across a typologically diverse range of landscape characters within a single study corridor. Third, MLLMs are better at identifying which elements are present in a landscape than at capturing how those elements are spatially arranged [16,21], consistent with evidence that vision–language models encode object-level semantics more reliably than the spatial relationships among objects [29]. The implications for landscape types whose value derives from holistic or contextual qualities remain largely unexplored. Finally, model breadth and prompt depth have never been crossed at scale. Recent urban-perception studies have compared up to six [28], seven [27], or nine [30] models, but the multi-model comparison was invariably conducted under a single scoring prompt and rating format, so that model differences cannot be separated from prompt sensitivity. Where prompt variation was examined at all, it was largely confined to a single model: prior-knowledge ablations were restricted to Claude-3.7-Sonnet [28], a role-play probe was applied only to GPT-4o [27], and three prompt formulations were tested on a single model [20]. In landscape preference research specifically, no more than two models have been compared [16,17], and even the one study to pair prompt adaptation with two models tested only a single characteristic-augmented variant, with the second model included merely as a sensitivity check [16]. Consequently, whether observed AI–human discrepancies reflect particular architectural limitations, prompt-design artefacts, or broader constraints of the MLLM paradigm cannot be disentangled by any existing design (Table 1). Beyond these methodological gaps, MLLM landscape preferences have yet to be systematically compared with those of road users in a tropical Southeast Asian corridor.

**Table 1.** Comparison of prior MLLM-based landscape and urban-perception preference studies with the present study across five methodological dimensions.

| Study | Models (*n*) | Prompt Formulations (n) | Evaluation Paradigm | Position-Bias Control | Landscape Scope |
|---|---|---|---|---|---|
| Lin et al. [16] | 2 GPT-4o, Qwen3 | 2 (baseline vs. characteristic-augmented "Apt") × 3 personas; Apt mainly on GPT-4o, Qwen3 as sensitivity check | Absolute rating | None | Agricultural heritage |
| Tung et al. [17] | 2 GPT-4, LLaVA | 1 (structured questionnaire) | Absolute rating | None | Natural scenes |
| Malekzadeh et al. [20] | 1 GPT-4 | 3 (progressively richer criteria) × 2 personas | Absolute rating | None | Urban street-level |
| Liu et al. [21] | 1 GPT-4 | 1 (BROKE role-play); 1333 agent personas | Absolute rating | None | Typical open spaces (urban/rural) |
| Zhou et al. [25] | 1 GPT-4o | 1 (no persona) | Pairwise; ternary | None reported | Urban street-level |
| Yoo et al. [26] | 1 GPT-4o | 1 (zero-shot, no persona) | Pairwise; ternary | No AB/BA | Urban street-level |
| Zhang et al. [27] | 7 5 open-source + GPT-4V, GPT-4o | 1 (expert persona); role-play tested on GPT-4o only | Pairwise; ternary | None reported | Urban street-level |
| Ma & Kwan [28] | 6 4 closed + 2 open | 1 (no persona); prior-knowledge ablation on Claude-3.7-Sonnet only | Absolute rating | None | Urban street-level |
| Yao et al. [30] | 9 6 open + 3 commercial | 2 role-based, but only 1 scores (2nd = renewal text, not in comparison) | Absolute rating | None | Urban street-level |

The objective of this study is to quantify how closely MLLM-derived visual preference rankings align with those of road users, to identify where and why the two diverge, and thereby to establish the conditions under which MLLMs can support corridor-scale visual assessment. This study addresses these gaps by conducting a systematic comparison of MLLM-derived and human visual preference rankings for highway landscape characters along a 418 km corridor of Malaysia's North–South Expressway. Five independently developed MLLMs were queried under three prompt formulations using a complete pairwise-comparison design with full AB/BA positional reversal, and their rankings were compared against ratings from 400 respondents across 16 landscape character groups. An additional forced-choice human survey ($N$ = 60) was administered to calibrate the response format. The study tested five hypotheses that (1) the five commercial MLLMs (ChatGPT, Claude, Gemini, Kimi, and Qwen) will produce convergent preference rankings for highway landscape images; (2) the AI ensemble's preference ranking will be affected by prompt formulation; (3) the AI-derived ensemble ranking will show a significant positive correlation with human preferences at both the image and group levels; (4) AI–human divergences will be systematic, concentrating in specific landscape character groups; and

(5) AI–human agreement will be affected by whether human and AI preferences are elicited using the same response format.

## 2. Materials and Methods

This study employed a mixed-method design comprising four stages (Figure 1). Stage 1 assembled the stimulus set of 80 georeferenced highway landscape images (five per group), drawn from a 1828-image corpus previously classified and validated in a companion study [31]. Stage 2 collected human visual preference ratings for all 80 images via a bilingual online survey ($N$ = 400); an independent pairwise validation survey ($N$ = 60) was additionally administered on a 48-image subset to calibrate the response format. Stage 3 conducted a complete pairwise comparison of all 3160 unique image pairs across five independently queried MLLMs under three prompt formulations, with full AB/BA positional reversal, yielding 6320 × 5 × 3 = 94,800 forced-choice judgements. Stage 4 integrated the human and AI preference data through Bradley–Terry modelling, cross-model reliability analysis, systematic AI–Human divergence identification, progressive prompt refinement, and instrument calibration. The following subsections detail each stage.

### *2.1. Study Area*

The study corridor comprises the 418.26 km northern section of Malaysia's North–South Expressway (NSE) (Figure 2), extending from Plaza Tol Jalan Duta in Kuala Lumpur northward to Alor Setar Utara Toll Plaza in Kedah. As the country's principal controlled-access highway, the NSE traverses multiple states and passes through a diverse mosaic of land-use contexts—urban peripheries, monoculture plantation belts, irrigated agricultural plains, and undulating forested terrain, thus providing substantial variation in roadside visual environments.

### *2.2. Image Acquisition and Stimulus Set Construction*

#### 2.2.1. Corridor-Scale Image Sampling

A preliminary vehicle-mounted survey using a dashboard camera was conducted along the entire corridor to establish georeferenced sampling points and map the spatial distribution of landscape characters. All images used in subsequent analyses were then sourced from Google Street View at the corresponding locations, as Street View provided standardised resolution, consistent viewing geometry, and reproducible capture conditions suitable for comparative preference assessment.

Images were extracted at approximately 250 m intervals along the northbound roadway. A quality-control procedure excluded frames affected by vehicular occlusion (e.g., preceding trucks blocking the forward view) or severely degraded atmospheric visibility. The resulting dataset comprised 1828 geo-referenced images (median spacing = 222 m; 90th percentile = 312 m), collectively documenting the visual environment of the entire corridor at a resolution sufficient to capture transitions between landscape characters.

#### 2.2.2. Landscape Character Classification

The visual environments along the study corridor were classified into 16 highway landscape character groups in a companion study [31], which documents the full classification protocol and its validation. Each character label captures the visually dominant elements in the driver's forward view. For example, a landscape character classified as "paddy and vegetation" may include some background hills, but the rice field is the defining character.

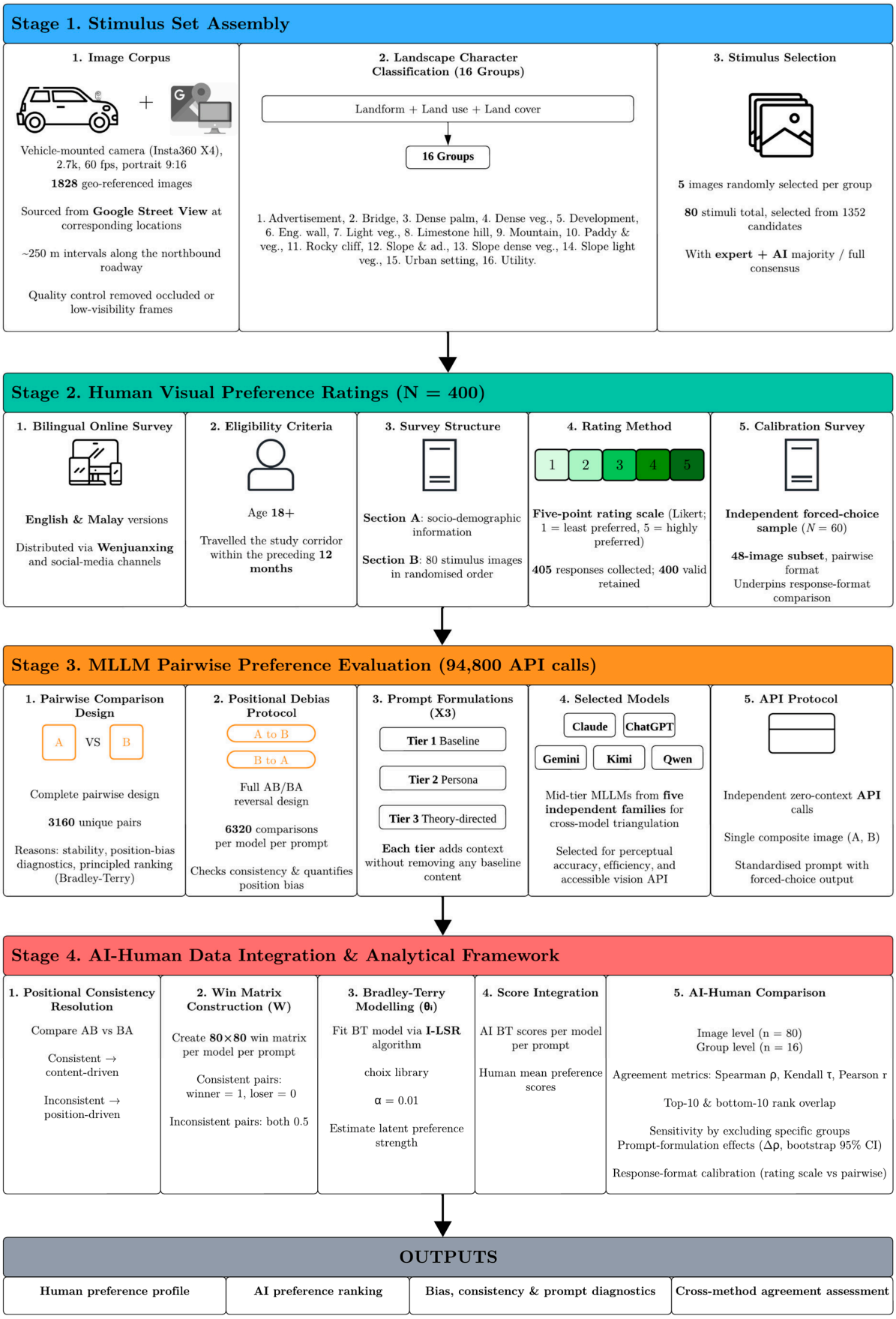


**Figure 1.** Overview of the four-stage research design: stimulus set assembly, human visual preference rating, MLLM pairwise preference evaluation, and AI–human data integration.


**Figure 2.** Location of the study corridor on the North–South Expressway (NSE) northern route, Peninsular Malaysia.

The 16 groups, ordered here from predominantly natural to predominantly anthropogenic, are: "mountain," "limestone hill," "rocky cliff," "dense vegetation," "slope dense vegetation," "slope light vegetation," "light vegetation," "dense palm plantation," "paddy and vegetation," "bridge," "engineered wall," "slope and advertisement," "advertisement," "development view," "utility," and "urban setting." Representative images are provided in Table S1 of the Supplementary Materials.

#### 2.2.3. Stimulus Selection

To construct the stimulus set for this study, five images were randomly selected per group, yielding a total of 80 images. Selection drew exclusively from the 1352 images that had achieved full or majority consensus in the companion study's triangulated validation—meaning at least two of three independently queried MLLMs agreed with the expert label [31]. The same 80-image set was used in both the human preference survey and the MLLM pairwise evaluation, enabling direct cross-method comparison on identical stimuli.

### *2.3. Human Preference Survey*

#### 2.3.1. Primary Likert Preference Survey

A bilingual online survey (English and Malay) was administered via Wenjuanxing (www.wjx.cn) and distributed through social media channels (WhatsApp, Facebook, Telegram, Instagram, WeChat) using convenience snowball sampling. Eligibility required respondents to be at least 18 years of age and to have travelled on the study corridor (Northern Route E1, Jalan Duta–Alor Setar Utara, in either direction) within the preceding

12 months. Both criteria were stated at the outset of the questionnaire and confirmed by respondents before proceeding.

The questionnaire comprised two sections. Section A collected socio-demographic and highway-use characteristics. Section B presented the 80 stimulus images described above, each displayed individually in a randomized order that varied across respondents to control for sequence effects. Visual preference denotes an immediate, holistic evaluative response to a viewed scene, reflecting the perceiver's rapid processing of environmental information rather than deliberate aesthetic judgement [32]. For each image, respondents rated their visual preference on a five-point rating scale (Likert; 1 = least preferred, 5 = highly preferred). An additional image was placed at the beginning and end of this section as familiarisation and buffer stimuli, so the ratings for these two images were excluded from the analysis. A total of 405 responses were collected, of which 400 passed completeness and quality checks and were retained for analysis.

#### 2.3.2. Pairwise Validation Survey

To place human and MLLM preferences on a common forced-choice footing and to test whether the response format itself shaped the elicited preferences, an independent pairwise validation survey was conducted. A subset of 48 of the 80 stimulus images was used, comprising three images randomly selected from the five available for each of the 16 landscape character groups. Sixty respondents—recruited separately from, and independent of, the 400 participants in the primary survey—each completed 50 forced-choice comparisons using the identical prompt applied to the MLLMs (Section 2.4.2). Across the sample, every one of the C (48, 2) = 1128 possible image pairs was evaluated at least once—7 pairs once, 370 twice, and 751 three times (mean = 2.7 judgements per pair)—so that the comparison graph was complete, with each image appearing in 119 to 130 comparisons (mean = 125). The 45 Malaysian citizens comprised 15 Malay, 15 Chinese, and 15 Indian respondents; the remaining 15 were non-citizens resident in Malaysia. Because this sample, image subset, and pairwise design are independent of the primary 80-image survey and of the MLLM C (80, 2) design, the two datasets are compared at the level of derived rankings rather than matched pair by pair.

### *2.4. MLLM Pairwise Visual Preference Evaluation*

To generate AI-based preference judgements, a complete pairwise comparison design was adopted rather than direct absolute ratings (e.g., asking each model to assign a 1–5 score to individual images).

Three considerations motivated the pairwise design for the AI evaluation. First, absolute single-answer grading tends to produce scores that fluctuate when the judging model changes, a form of instability less pronounced in pairwise comparison [23], making direct comparison of absolute ratings across models or against human raters problematic. Pairwise forced-choice judgements circumvent this calibration problem by requiring only ordinal discrimination ("which is better?") rather than cardinal magnitude estimation ("how good?"). This advantage has also been empirically demonstrated in subjective image quality assessment, where forced-choice pairwise comparison yields smaller measurement variance than rating-based methods [33]. Second, pairwise designs allow systematic detection and quantification of position bias through AB/BA reversal, a diagnostic capability that absolute rating cannot provide [24]. Third, the Bradley–Terry model offers a principled statistical framework for recovering a complete preference ranking from pairwise data [34], producing continuous-scale scores that can be directly correlated with human-derived mean ratings.

#### 2.4.1. Pairwise Design and Positional Debiasing

Each image pair was presented within each model in both presentation orders: the original order (AB) and the reversed order (BA), following the positional debiasing protocol proposed by Shi et al. [24]. This full AB/BA reversal design doubled the number of comparisons per model from 3160 to 6320, yielding a total of 6320 × 5 models × 3 prompt formulations = 94,800 API calls.

#### 2.4.2. API Protocol

Each comparison was submitted as an independent, zero-context API call, with the decoding temperature set to 0. The model received a single composite image containing the two images side by side (labelled A and B), a standardised prompt, and no prior conversational history. The baseline prompt (Tier 1; Section 2.4.4) was identical across all five models and both presentation orders:

"You are comparing two highway landscape scenes along Malaysia's North–South Expressway. Which scene do you find more visually preferable? Focus only on scenic quality. Ignore weather, traffic, and road engineering. Respond with only: A or B."

The instruction to respond with only "A" or "B" constrained the output to a forced-choice format, eliminating hedging, ties, or extended reasoning that could introduce parsing ambiguity. No follow-up prompts, chain-of-thought elicitation, or iterative feedback were employed. The zero-context design ensured that each judgement was independent, preventing carry-over effects from preceding comparisons.

#### 2.4.3. Model Selection Rationale

Five models were selected from distinct commercial MLLM families: Claude Sonnet 4.6 (Anthropic, San Francisco, CA, USA), ChatGPT 5.4 (OpenAI, San Francisco, CA, USA), Gemini 3 Flash Preview (Google, Mountain View, CA, USA), Kimi K2.5 (Moonshot AI, Beijing, China), and Qwen3.5-Plus (Alibaba, Hangzhou, China).

First, the pairwise comparison task is a visual discrimination task with a constrained binary output, requiring neither multi-step reasoning nor contextual memory. These are precisely the capabilities that distinguish flagship models from their mid-tier counterparts: recent benchmark evidence indicates that the mid-tier-to-flagship performance gap is concentrated in higher-order cognitive reasoning, whereas perceptual accuracy remains largely saturated across parameter scales [22]. For a purely perceptual task, mid-tier models therefore match flagship-level performance while remaining economically viable at scale (94,800 API calls across the five-model, three-prompt design).

Second, lightweight models were excluded because fine-grained visual recognition remains challenging even for full-scale MLLMs [35], and further reducing model capacity risks compromising the discrimination needed to distinguish visually similar landscape character groups (e.g., "slope dense vegetation" from "dense vegetation").

#### 2.4.4. Progressive Prompt Refinement

The three-tier design follows a general principle in prompting research: progressively enriching the task context isolates whether a model's errors stem from limited capability or from limited attention to task-relevant cues. Holding the stimulus set and protocol fixed while adding only contextual framing turns prompt formulation into a controlled manipulation, so that any change in output is attributable to the added information rather than to the model or the items. The specific progression—from a neutral baseline, to a role-based persona, to explicit theory-derived criteria—spans the two levers most commonly used to steer LLM judgement: perspective-taking (persona prompting) and criterion specification

(instruction grounding). This progression generalises to any evaluative task that requires separating what a model can perceive from what it weighs.

Each tier was administered under the identical pairwise protocol, positional reversal, and zero-context conditions (Sections 2.4.1 and 2.4.2), and each added information without removing any. Tier 1 was the baseline instruction, with no contextual framing. Tier 2 prepended a road-user persona, instructing the model to respond as a person who had travelled the northern NSE corridor at least once in the preceding year, whether driving or as a passenger, thereby testing whether adopting a local road-user perspective alters model preferences. Tier 3 retained this persona and additionally directed attention to two perceptual dimensions drawn from established landscape-preference theory: (i) the openness of the outlook—whether the eye can survey into the distance while the scene still affords some sense of enclosure or shelter—following prospect–refuge theory [36]; and (ii) the naturalness of the scene, drawing on attention restoration theory [32,37]: natural content tends to be restorative and visually preferred, while conspicuous human-made elements that fail to blend into the scene tend to lower perceived scenic quality [38]. Crucially, no landscape character group was named or alluded to in any of the three prompts; Tier 3 supplied only general, theory-derived perceptual criteria, so that any resulting change in the ranking reflects redirected attention rather than an instruction to favour a particular landscape character group. The full wording of all three prompts is provided in Appendix A.

### *2.5. Analytical Framework*

All analyses were carried out within a single, unified pipeline. Because the same five MLLMs were queried under three prompt formulations of increasing guidance (Tiers 1–3; Section 2.4.4), the entire pipeline below was run once per formulation: every step, from consistency resolution and win-matrix construction through Bradley–Terry estimation, cross-model agreement, the AI preference ranking, and the AI–human comparison (Sections 2.5.1–2.5.6), was repeated under each condition. The instrument calibration in Section 2.5.7 is the single exception; as a one-off check on the human response format, it is independent of the prompt formulations and was performed once.

#### 2.5.1. Positional Consistency Resolution

For each of the 3160 pairs, the AB and BA results were compared to determine whether the model consistently preferred the same image regardless of presentation order. A pair was classified as consistent when the winning image was the same across both rounds, and inconsistent when the two rounds preferred different images—the latter indicating that the judgement was driven by position rather than content. This resolution was performed independently for each model and prompt formulation, yielding a per-model consistency rate and a characterisation of each model's bias direction (primacy vs. recency). Bias direction was assessed among the inconsistent pairs only: for each such pair, selecting the image shown first in both rounds indicates a primacy effect and selecting the image shown second a recency effect, and the proportion of a model's inconsistent pairs falling into each category characterised the direction—rather than merely the rate—of its position bias.

#### 2.5.2. Win Matrix Construction

For each model under each prompt formulation, a complete $80 \times 80$ win matrix W was constructed. For consistent pairs, the winner received W[i][j] = 1 and the loser W[j][i] = 0. For inconsistent pairs (positional ties), both entries were set to 0.5, treating the outcome as uninformative. Because the design covers all C (80, 2) pairs, every off-diagonal cell in W is filled, guaranteeing that the subsequent Bradley–Terry estimation is well-defined.

To derive the ensemble preference ranking within each prompt formulation, the five models' consistency-resolved winners (from Section 2.5.1) were aggregated by majority vote across all 3160 pairs. For each pair, only the models that produced a consistent judgement contributed a vote; the image preferred by the larger number of these consistent votes was recorded as the ensemble winner. Pairs on which the consistent votes were evenly split, or on which no model produced a consistent judgement, were coded as ensemble ties (W[i][j] = W[j][i] = 0.5). For example, if four of the five models produced a consistent judgement for a pair, a three-to-one split among them yielded an ensemble winner while a two-to-two split was coded as a tie; the fifth (inconsistent) model contributed no vote in either case. An ensemble win matrix was then constructed and fitted using the same Bradley–Terry procedure described in Section 2.5.3.

#### 2.5.3. Bradley–Terry Modelling

The Bradley–Terry model [34] was fitted to each model's win matrix using the Iterative Luce Spectral Ranking (I-LSR) algorithm [39], as implemented in the Python 3.10 library choix (https://github.com/lucasmaystre/choix, (accessed on 2 April 2026)), with a smoothing parameter $\alpha = 0.01$ to ensure numerical stability on sparse comparison graphs. The BT model estimates a latent preference-strength parameter $\theta_i$ for each image, such that the probability of image i being preferred over image j is given by $\sigma(\theta_i - \theta_j)$, where $\sigma(\cdot)$ is the logistic function. Fitted once per model–prompt combination and once per prompt-level ensemble, this procedure produced, for each prompt formulation, five model rankings and one ensemble ranking over the 80 images.

#### 2.5.4. Cross-Model Agreement

Within each prompt formulation, agreement among the five model rankings was quantified at the image level. Pairwise Spearman rank correlations ($\rho$) and Kendall rank correlations ($\tau$) were computed between every pair of models' Bradley–Terry rankings, and Kendall's coefficient of concordance (W) was computed across all five rankings to summarise overall convergence. Pair-level agreement was further described as the proportion of the 3160 image pairs on which the models returned a unanimous verdict and the proportion on which they reached a majority verdict. The internal coherence of each model's preferences was additionally checked for transitivity: among the image triples for which all three pairwise comparisons yielded a strict (non-tied) winner, the proportion forming an intransitive (cyclic) preference loop was computed, with low values indicating that the transitivity assumption underlying the Bradley–Terry model is satisfied. This step verifies that the ensemble ranking is sufficiently representative to serve as the AI benchmark in the human comparison (Section 2.5.6).

#### 2.5.5. AI Preference Ranking

For each prompt formulation, the ensemble Bradley–Terry ranking (Sections 2.5.2 and 2.5.3) was summarised to characterise the models' collective landscape preferences. Image-level preference strengths were aggregated at the group level by averaging the five images within each group, and the resulting group scores were min–max rescaled to the [0, 1] interval for each model and the ensemble separately, to aid interpretation. Because the cross-model agreement established in Section 2.5.4 confirms that this ensemble is representative of the individual models, the group-level ensemble ordering serves as the consensus AI preference that enters the human comparison. Repeating this across the three formulations shows how the models' collective preference ordering shifts as prompt guidance increases.

#### 2.5.6. AI–Human Comparison

Within each prompt formulation, the BT scores (per-model and ensemble) were compared with the human mean preference scores ($N$ = 400) at two levels of granularity: image level ($n$ = 80) and group level ($n$ = 16, computed as the mean score of the five images within each group). Agreement was quantified using Spearman's rank correlation (ρ) and Kendall's rank correlation (τ) at both the image and group levels, with Pearson's linear correlation (r) additionally reported at the image level. Top-10 and bottom-10 rank overlap was computed to assess convergence at the extremes of the preference distribution. To quantify the uncertainty of these correlations and to test whether AI–human agreement changed across prompt formulations, 95% confidence intervals were obtained by bootstrap resampling (5000 resamples of the 80 images at the image level, or the 16 groups at the group level), and the Tier 1-to-Tier 3 difference in Spearman's ρ was tested by resampling the same units and computing the bootstrap distribution of the difference. Because this comparison was repeated across all three formulations, the ranks of the 16 groups were also tracked across formulations to identify those whose alignment with human preference shifted as guidance increased. To quantify how much of the divergence was driven by individual groups, the image- and group-level correlations were recomputed, excluding the paddy-and-vegetation group (its five images).

#### 2.5.7. Instrument Calibration: Likert Versus Pairwise

To evaluate whether the response-format asymmetry between the human (rating scale) and MLLM (forced-choice) tasks could account for the AI–human divergences, the human pairwise judgements from the validation survey (Section 2.3.2) were aggregated into a win matrix—each forced choice contributing one vote—and fitted with the same Bradley–Terry procedure (Section 2.5.3), yielding a human pairwise preference ranking over the 48-image subset. Two comparisons were then computed at the image level ($n$ = 48): (i) the human pairwise ranking against the Likert-derived human ranking from the primary survey, quantifying the convergence of the two human instruments; and (ii) the AI ensemble ranking against the human pairwise ranking, providing a same-paradigm (forced-choice) AI–human comparison alongside the cross-paradigm comparison of Section 2.5.6. All correlations were estimated using Spearman's ρ, Kendall's τ, and Pearson's r, with 95% bootstrap confidence intervals (5000 resamples of the 48 images). A close correspondence between the rating scale and pairwise human rankings, together with higher AI–human agreement under a common paradigm, would indicate that the AI–human divergences reported in Section 3.6 cannot be attributed solely to response-format asymmetry.

## 3. Results

### *3.1. Respondent Profile*

The demographic and highway-use profile of the 400 valid respondents is summarised in Table S2. Respondents were near-evenly split by gender (47.8% male, 52.2% female), concentrated in middle adulthood (60.0% aged 25–44), and predominantly Malaysian citizens (86.3%), with an ethnic composition broadly reflecting national diversity (Malay 41.8%, Chinese 34.0%, Indian 13.0%). Over half held a bachelor's degree or above (52.8%), and half resided in the Klang Valley conurbation (50.3%). Monthly household income was spread across brackets from lower-middle to upper-middle. The vast majority held driving licences (87.8%), and all respondents had travelled the study corridor at least once in the preceding 12 months, with most (62.5%) having done so one to three times, confirming first-hand familiarity with the roadside visual environment. Nearly three-quarters (73.0%) described the corridor's landscapes as interesting, indicating pre-existing engagement with roadside scenery.

### *3.2. Human Preference Baseline*

Across the 80 images, human mean preference scores ranged from 2.163 to 3.607 on the five-point scale (grand mean = 2.936, SD = 0.417). The distribution was approximately normal (Shapiro–Wilk W = 0.994, $p = 0.159$), and inter-rater agreement was high (Cronbach's $\alpha = 0.967$), confirming that the aggregated scores provide a stable baseline for comparison with AI-derived rankings. Individual image scores are reported in Table S3 of the Supplementary Materials.

At the group level (Table 2), group means ranged from 2.250 ("urban setting") to 3.579 ("paddy and vegetation"). Three preference tiers are discernible. The highest-rated tier—"paddy and vegetation" (3.579), "mountain" (3.455), and "limestone hill" (3.350)—represents landscapes dominated by natural landforms or agrarian openness. A middle tier spanning "dense vegetation" (3.282) down to "dense palm plantation" (2.910) comprises vegetated and mixed characters. The lowest-rated tier—from "bridge" (2.773) to "urban setting" (2.250)—is dominated by built infrastructure. This human preference hierarchy serves as the benchmark against which AI-derived rankings are evaluated in the following sections.

**Table 2.** Ranked group-mean preference scores for 16 highway landscape character groups (five photos per group; $N = 400$).

| Rank | Landscape Character | Group Mean | Photo IDs (5) |
|---|---|---|---|
| 1 | Paddy and vegetation | 3.579 | 25, 29, 50, 73, 77 |
| 2 | Mountain | 3.455 | 13, 17, 40, 53, 55 |
| 3 | Limestone hill | 3.350 | 2, 4, 45, 79, 80 |
| 4 | Dense vegetation | 3.282 | 21, 37, 38, 51, 54 |
| 5 | Rocky cliff | 3.201 | 27, 31, 47, 66, 67 |
| 6 | Slope dense vegetation | 3.130 | 16, 22, 61, 72, 78 |
| 7 | Slope light vegetation | 3.061 | 1, 42, 57, 68, 74 |
| 8 | Engineered wall | 3.028 | 6, 10, 14, 23, 64 |
| 9 | Light vegetation | 2.940 | 18, 19, 33, 35, 56 |
| 10 | Dense palm plantation | 2.910 | 8, 44, 60, 75, 76 |
| 11 | Bridge | 2.773 | 26, 36, 63, 69, 70 |
| 12 | Slope and advertisement | 2.645 | 5, 39, 43, 52, 59 |
| 13 | Advertisement | 2.573 | 11, 30, 32, 48, 71 |
| 14 | Utility | 2.458 | 7, 9, 28, 34, 65 |
| 15 | Development view | 2.344 | 12, 20, 24, 46, 62 |
| 16 | Urban setting | 2.250 | 3, 15, 41, 49, 58 |

### *3.3. MLLM Positional Consistency and Bias*

Before testing the hypotheses, this section establishes the positional reliability of the five models on which the subsequent hypothesis tests depend. The AB/BA reversal design was applied to all five models under each of the three prompt formulations (Figure 3). Two patterns emerged: the rate of positional consistency was a stable, model-specific property, whereas the direction of the residual bias was less robust and, for some models, sensitive to prompt wording.

Consistency itself was strongly model-specific yet largely invariant to prompt wording. Under the baseline prompt, ChatGPT was the most reliable, returning the same winner regardless of presentation order on 93.3% of the 3160 pairs, followed by Qwen (86.0%), Gemini (85.9%), Kimi (84.1%), and Claude (80.5%). This ordering—ChatGPT clearly highest, Claude lowest, and the other three closely clustered—was preserved under the persona and theory-directed prompts, and no model's consistency shifted by more than about four percentage points across the three tiers (Figure 3). Positional reliability is therefore

an intrinsic property of each model rather than an artefact of prompt wording: richer prompting neither degraded nor systematically improved robustness to presentation order.

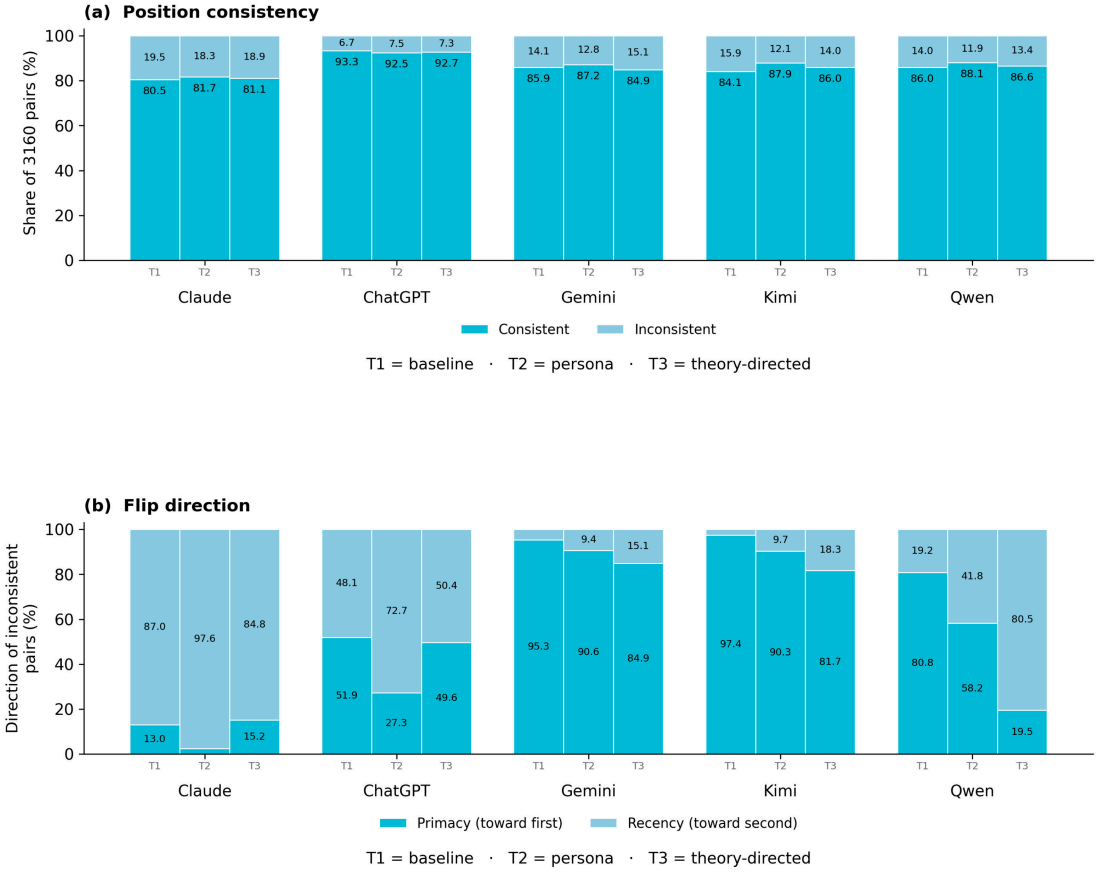


**Figure 3.** AB/BA positional reliability of the five MLLMs under the three prompt formulations (Tier 1 baseline, Tier 2 persona, Tier 3 theory-directed), based on the 3160 paired image comparisons per model and condition.

The direction of these residual flips, by contrast, was far less robust than the consistency rate. Under the baseline prompt, the models fell into three groups: Gemini (95.3% of its 445 inconsistent pairs), Kimi (97.4% of 502), and Qwen (80.8% of 443) flipped predominantly toward the first-presented image (a primacy bias); Claude flipped toward the second (a recency bias, 87.0% of 617 pairs); and ChatGPT's flips were almost evenly split (51.9% versus 48.1%, i.e., no directional preference). Across the persona- and theory-directed prompts, three tendencies held: Claude remained recency-biased, and Gemini and Kimi remained primacy-biased under all three formulations. The other two did not. Qwen reversed cleanly from an 80.8% primacy bias at baseline to an 80.5% recency bias under the theory-directed prompt, and ChatGPT, near-balanced throughout, drifted to a 72.7% recency tendency under the persona prompt before returning to balance. The full primacy/recency breakdown for all five models across the three formulations is reported in Table S4 of the Supplementary Materials.

These inconsistency rates determine the information content of the win matrices, since each inconsistent pair contributes two uninformative tie cells (W[i][j] = W[j][i] = 0.5). At baseline, ChatGPT's matrix contained the fewest ties (424 off-diagonal cells, 6.7%), followed by Qwen (886), Gemini (890), Kimi (1004), and Claude (1234, 19.5%). ChatGPT's Bradley–Terry scores therefore have the highest discriminative resolution, and Claude's the lowest, a difference that propagates into the per-model rankings analysed below.

### *3.4. Cross-Model Agreement*

This section tests H1, which predicted convergent preference rankings across the five independently developed MLLMs. Having established that each model's rankings are internally reliable, we next assessed how closely the five models agreed with one another. Agreement was high under all three prompt formulations and, although it weakened slightly as prompt guidance increased, it remained strong throughout. Across the five models' Bradley–Terry rankings, the mean pairwise Spearman correlation was ρ = 0.908 under the baseline prompt (range 0.860–0.953), ρ = 0.878 under the persona prompt (0.803–0.952), and ρ = 0.886 under the theory-directed prompt (0.836–0.948); the overall five-model concordance followed the same ordering (Kendall's W = 0.926, 0.902, and 0.909, respectively; Figure 4). Pairwise Kendall's τ told the same story in all three conditions (mean τ = 0.750, 0.710, and 0.727; full values in Table S5). That the baseline showed the tightest agreement of the three indicates the added guidance gave the models marginally more room to diverge rather than less, but the effect was small—no formulation's mean correlation fell below ρ = 0.87.

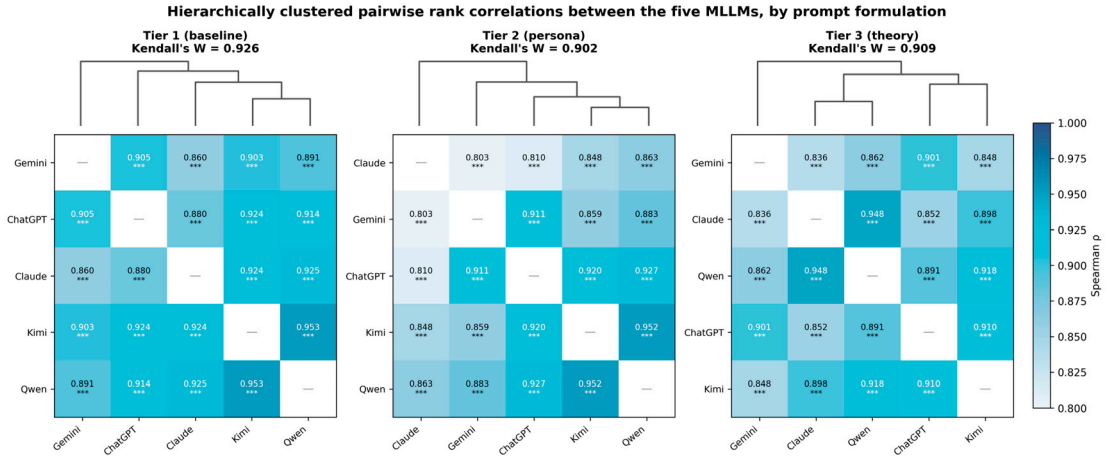


**Figure 4.** Hierarchically clustered pairwise rank correlations (Spearman's ρ) between the five MLLMs' Bradley–Terry rankings over the 80 images, for each prompt formulation. Within each panel, models are ordered by hierarchical clustering of their correlation distance (1 − ρ; dendrogram above), and cells are annotated with ρ and its significance. Kendall's coefficient of concordance (W) for each formulation appears in the panel title. *** $p < 0.001$ (all model pairs, $n = 80$).

At the level of individual image pairs, the same picture held under every formulation (Table 3): all five models returned the same winner on 55.6% of pairs under the baseline prompt, 52.5% under the persona prompt, and 56.0% under the theory-directed prompt, and a majority winner (at least four of five models agreeing) on 73.2%, 72.5%, and 71.7%, respectively. On the pairs where at least one model produced a consistent verdict (3153–3156 of the 3160), unanimous agreement held for roughly 53–56% and majority agreement for roughly 72–73%. That these proportions were reproduced under each of the three independently varied prompts confirms that the convergence is a property of the models rather than of any single elicitation.

These convergent rankings were also internally coherent under all three formulations. For each model, the win matrix was screened for intransitive (cyclic) triples—image triples whose pairwise outcomes formed a preference loop (i > j, j > k, k > i)—among the triples for which all three pairwise comparisons produced a strict (non-tied) winner. Such cycles were rare throughout: the proportion of strictly decided triples that were cyclic ranged from 0.018% to 0.140% across the five models under the baseline prompt, from 0.038% to 0.182% under the persona prompt, and from 0.026% to 0.205% under the theory-directed prompt.

The single departure from this pattern was Claude under the persona prompt (1.168%)—the same condition in which Claude showed its strongest recency bias (Section 3.3) and its lowest cross-model correlations—yet even there, 98.8% of decided triples were transitive. The full breakdown across the five models and three formulations is reported in Table S6. Intransitivity at this level confirms that the transitive-preference assumption underlying the Bradley–Terry model is well satisfied across all prompts. Hence, the latent preference strengths reported below are not artefacts of cyclic inconsistency.

**Table 3.** Pair-level agreement among the five MLLMs over the 3160 image pairs, under each prompt formulation.

| Prompt Formulation | Pairs Evaluated | Unanimous Agreement (5 of 5) | Majority Agreement (≥4 of 5) |
|---|---|---|---|
| Tier 1 (baseline) | 3153 | 1754 (55.6%) | 2308 (73.2%) |
| Tier 2 (persona) | 3155 | 1656 (52.5%) | 2287 (72.5%) |
| Tier 3 (theory) | 3156 | 1768 (56.0%) | 2262 (71.7%) |

Pairs evaluated = the number of pairs on which at least one model produced a consistent verdict, and percentages are computed on this basis. Pairs on which all five models reversed their choice under AB/BA presentation (and therefore cast no vote) are excluded: 7, 5, and 4 such pairs under the baseline, persona, and theory-directed prompts, respectively, giving 3153, 3155, and 3156 of the 3160 pairs.

The agreement was also broadly uniform across models rather than driven by a single pair, in every formulation. Kimi and Qwen were the most similar models throughout; their correlation never fell below $\rho = 0.918$ ($\rho = 0.953$, 0.952, and 0.918 under the baseline, persona, and theory-directed prompts), and they formed the strongest pair under the baseline and persona prompts. Claude was consistently the most distinctive: its correlations with the other models were the lowest in each condition, and they fell furthest under the persona prompt ($\rho = 0.803$–$0.863$), consistent with its comparatively strong recency tendency in that condition (Section 3.3). Even at this widest separation, no pairwise correlation fell below $\rho = 0.80$ in any of the three formulations (the minimum was $\rho = 0.803$, between Claude and Gemini under the persona prompt). The ensemble Bradley–Terry ranking, therefore, rests on genuine consensus across all five models for each prompt, rather than agreement among a subset, and this provides the basis for treating it as a meaningful summary of MLLM preferences in the human comparison that follows (Section 3.6).

### *3.5. AI Preference Ranking: Image and Group Level*

This section tests H2, which predicted that the AI ensemble's preference ranking would be affected by prompt formulation. At the image level, the five-model ensemble Bradley–Terry scores for all 80 images under the three prompt formulations are reported in Table S7. The extremes of the ranking were stable across all three prompts: the five highest-ranked images were mountain or limestone-hill characters under every formulation (IMG_79, IMG_55, IMG_04, IMG_45, and IMG_40 occupied the top five at baseline, with only minor reordering under the persona and theory-directed prompts), and the five lowest-ranked images were consistently urban-setting, utility, and development-view characters (IMG_03, IMG_58, IMG_09, IMG_46, and IMG_49 anchored the bottom under all three). Geomorphologically distinctive landforms thus dominated the top of the ranking, while built and utility characters clustered at the bottom, regardless of prompt wording.

The middle of the image-level ranking, by contrast, behaved exactly as described in the group-level analysis below. The baseline and persona rankings were closely similar—most images moved only a few positions between them—whereas the theory-directed prompt substantially reordered the central images. The clearest image-level movements under the theory-directed prompt were a coherent rise of paddy-and-vegetation characters and a fall of advertisement and some engineered characters. All five paddy-and-vegetation images

rose (for example, IMG_50 from rank 58 to 22, IMG_73 from 60 to 26, and IMG_29 from 25 to 12), together with slope vegetation and palm plantation characters (IMG_22 from 47 to 13; IMG_76 from 36 to 14), while advertisement characters (IMG_43 from 13 to 38; IMG_05 from 28 to 47) and two rocky cliff characters (IMG_31 from 9 to 39; IMG_27 from 11 to 28) fell. These group-level shifts constitute the character-level reordering quantified in the following paragraphs. Aggregating to the group level (Table 4; Figure 5), the baseline ensemble means ranged from 0.943 ("mountain") to 0.179 ("urban setting"). The ordering was anchored at both ends: a high-preference group of geomorphologically distinctive characters ("mountain," "limestone hill," "rocky cliff"; BT $\geq$ 0.84) and a low-preference group of built and utility characters ("development view," "utility," "urban setting"; BT $\leq$ 0.41), with a broad middle band of vegetation, infrastructure, paddy, and advertisement characters (BT $\approx$ 0.53–0.79) in between. This baseline ordering was highly consistent across the five individual models (pairwise Kendall's $\tau$ = 0.72–0.90 at the group level), confirming that the ensemble represents a shared preference rather than the view of any single model.

**Table 4.** Mean ensemble Bradley–Terry scores for the 16 landscape character groups under each prompt formulation, with the rank of each group within that formulation in parentheses.

| Landscape Character | Tier 1 (Baseline) | Tier 2 (Persona) | Tier 3 (Theory) |
|---|---|---|---|
| Mountain | 0.943 (1) | 0.946 (1) | 0.915 (2) |
| Limestone hill | 0.913 (2) | 0.928 (2) | 0.927 (1) |
| Rocky cliff | 0.848 (3) | 0.840 (3) | 0.778 (6) |
| Dense vegetation | 0.789 (4) | 0.793 (4) | 0.779 (5) |
| Slope dense vegetation | 0.716 (5) | 0.723 (5) | 0.793 (4) |
| Slope light vegetation | 0.706 (6) | 0.716 (6) | 0.774 (7) |
| Slope and advertisement | 0.695 (7) | 0.697 (7) | 0.608 (10) |
| Light vegetation | 0.647 (8) | 0.639 (10) | 0.738 (9) |
| Advertisement | 0.644 (9) | 0.661 (8) | 0.579 (11) |
| Dense palm plantation | 0.622 (10) | 0.642 (9) | 0.743 (8) |
| Engineered wall | 0.613 (11) | 0.614 (11) | 0.566 (12) |
| Bridge | 0.592 (12) | 0.597 (12) | 0.459 (13) |
| Paddy and vegetation | 0.535 (13) | 0.570 (13) | 0.806 (3) |
| Development view | 0.404 (14) | 0.392 (14) | 0.387 (14) |
| Utility | 0.321 (15) | 0.324 (15) | 0.369 (15) |
| Urban setting | 0.179 (16) | 0.179 (16) | 0.122 (16) |

Repeating the ranking under the three prompt formulations showed that this AI preference ordering was not fixed but depended on the level of guidance, and did so in a specific way (Table 4; Figure 6). The baseline and persona rankings were almost identical (ensemble-level Kendall's $\tau$ = 0.967): adopting a road-user persona left the models' collective ordering essentially unchanged. The theory-directed prompt, by contrast, substantially reordered the middle of the ranking ($\tau$ = 0.700 against both the baseline and the persona). Under this prompt, paddy and vegetation characters rose, while advertisement-dominated characters declined. Some engineered characters moved up into the upper-middle ranks: "slope dense vegetation," "light vegetation," and "dense palm plantation"; "advertisement" and "slope and advertisement" dropped toward the lower-middle; and "rocky cliff" slipped from third to sixth. The largest single movement was "paddy and vegetation," which rose from rank 13 under both the baseline and persona prompts to rank 3 under the theory-directed prompt; this shift was consistent across all five models, each of which placed paddy between ranks 9 and 14 under the first two prompts and between ranks 3 and 8 under the theory-directed prompt. The high- and low-preference anchors, however, were unaffected by prompt wording: mountain and limestone hill remained the two top-ranked

characters, and urban setting, utility, and development views remained the three lowest, across all formulations. These three ensemble rankings constitute the AI-side input for the comparison against human preference in Section 3.6.

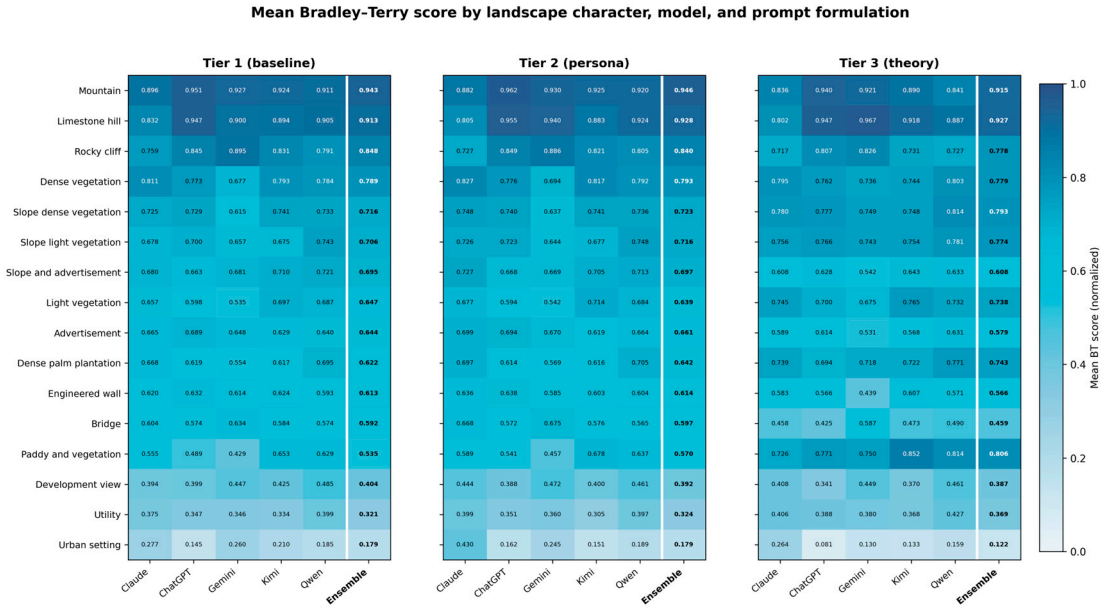


**Figure 5.** Mean Bradley–Terry scores for the 16 landscape character groups, shown for each of the five MLLMs and their majority-vote ensemble (bold, rightmost column of each panel), under the three prompt formulations (Tier 1 baseline, Tier 2 persona, Tier 3 theory-directed).

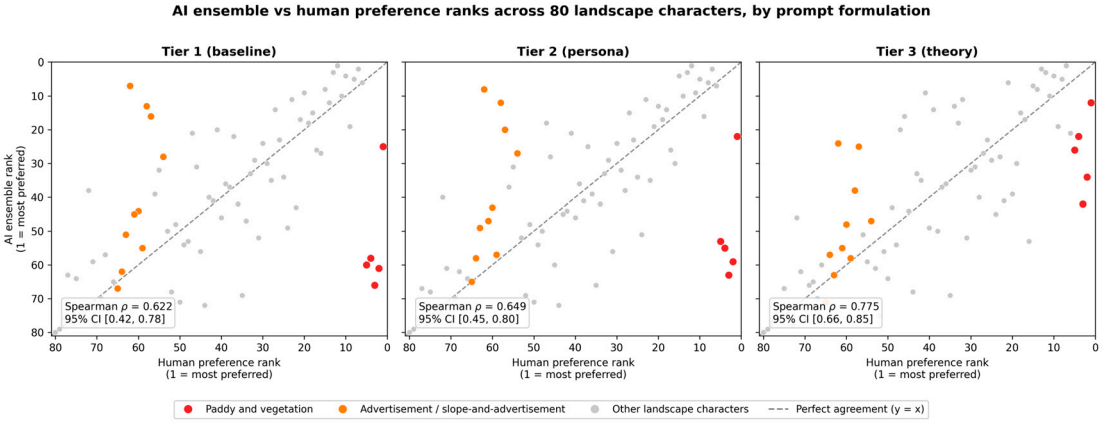


**Figure 6.** AI ensemble versus human preference ranks for the 80 images, by prompt formulation (Tier 1 baseline, Tier 2 persona, Tier 3 theory-directed).

### *3.6. AI–Human Comparison*

This section tests H3 and H4, which predicted, respectively, a significant positive correlation between the AI ensemble ranking and human preferences, and systematic AI–human divergences concentrated in specific landscape character groups. We compared the AI ensemble rankings with the human preference benchmark ($N$ = 400) at both the image level (80 images) and the group level (16 landscape character groups; Table 5). Under the baseline prompt, agreement was moderate rather than strong: the AI ensemble and human rankings correlated at ρ = 0.622 (95% CI [0.42, 0.78]) across the 80 images and ρ = 0.697 across the 16 groups, with concordant Kendall's τ and Pearson's r (Table 5). The two rankings thus shared a broad ordering—both placed natural and geomorphologically distinctive images above built and utility ones—yet diverged substantially on specific groups (Figure 6).

**Table 5.** Agreement between AI and human landscape-preference rankings under each prompt formulation, at the image level (80 images) and the group level (16 groups), with top-10 and bottom-10 image overlap.

| Prompt Formulation | Image Level ($n$ = 80) | | | Group Level ($n$ = 16) | | Top-10/Bottom-10 Overlap |
|---|---|---|---|---|---|---|
| | Spearman's ρ [95% CI] | τ | r | Spearman's ρ [95% CI] | τ | (Images) |
| Tier 1 (baseline) | 0.622 [0.42, 0.78] | 0.482 | 0.648 | 0.697 [0.18, 0.98] | 0.617 | 4/6 |
| Tier 2 (persona) | 0.649 [0.45, 0.80] | 0.505 | 0.671 | 0.682 [0.14, 0.98] | 0.583 | 4/6 |
| Tier 3 (theory) | 0.775 [0.66, 0.86] | 0.586 | 0.798 | 0.929 [0.74, 0.98] | 0.783 | 3/7 |

This agreement was not fixed across prompt formulations. Adopting a road-user persona (Tier 2) left it essentially unchanged (image ρ = 0.649; group ρ = 0.682), but the theory-directed prompt (Tier 3) raised it markedly, to ρ = 0.775 (95% CI [0.66, 0.86]) at the image level and ρ = 0.929 at the group level. At the image level, where the 80-image sample affords adequate statistical power, this improvement was significant: the Tier 1-to-Tier 3 gain of Δρ = +0.153 had a bootstrap 95% CI of [0.029, 0.289] (one-sided $p$ = 0.007), and the gain over the persona prompt was likewise significant (Δρ = +0.125, 95% CI [0.012, 0.250]). The group-level point estimate rose in parallel (0.697 to 0.929), corroborating the image-level result; however, with only 16 groups, the group-level confidence intervals are wide (e.g., [0.18, 0.98] at baseline), and the Tier 1-to-Tier 3 difference did not itself reach significance (Δρ = +0.232, 95% CI [−0.045, +0.707]). We therefore treat the group-level correlations as descriptive and rest the inferential claim on the image level: theory-directed prompting brought the models' preferences significantly closer to human preference, whereas persona framing did not.

The source of both the baseline divergence and its correction is evident at the group level (Figure 6; Table 4). The largest baseline disagreement concerned the paddy-and-vegetation group, which human raters ranked highest (rank 1) among the 16 groups but which the AI ensemble ranked near the bottom (rank 13) under both the baseline and persona prompts—a 12-position undervaluation. AI also mildly overvalued the advertisement and slope-and-advertisement groups relative to human raters, over-ranking them by four to five positions. Under the theory-directed prompt, both gaps closed: the paddy-and-vegetation group rose to rank 3, its five constituent images each moving from the lower-middle into the upper portion of the ranking (Figure 6; Table S7), while the advertisement groups fell to the lower-middle ranks where human raters had placed them. After this correction, the only group-level discrepancy of four or more positions remaining under Tier 3 was a modest under-ranking of the engineered-wall group. These group-level shifts across the three prompt formulations are shown in Figure 7.

Agreement at the extremes of the distribution followed the same pattern less sharply. The AI and human rankings shared 6 of their bottom-10 images under the baseline prompt and 7 under the theory-directed prompt—both consistently identifying the urban, utility, and development groups as least preferred—whereas top-10 overlap remained modest throughout (3–4 of 10), because human raters placed the paddy group at the very top while the AI ranking, even after correction, was still led by the mountain and limestone-hill groups. Taken together, the comparison shows that the MLLM ensemble reproduces the broad shape of human landscape preference; that its principal systematic departure—the undervaluation of cultivated paddy landscape—is specific and group-bound; and that this departure is substantially corrected by theory-directed prompting, although the exact ordering of the most-preferred images continues to differ. To gauge how much of the AI–human divergence was attributable to this single group, agreement was recomputed after removing the five paddy-and-vegetation images (Figure 8). Under the baseline prompt,

excluding paddy-raised image-level agreement from ρ = 0.622 to 0.772 and group-level agreement from ρ = 0.697 to 0.911, the persona prompt behaved similarly (image: 0.649 to 0.780; group: 0.682 to 0.893). Under the theory-directed prompt, by contrast, the exclusion changed agreement only marginally (image 0.775 to 0.801; group 0.929 to 0.925), because that prompt had already corrected the paddy undervaluation. Two observations follow. First, a single landscape character group accounted for the bulk of the baseline divergence: with paddy set aside, the ensemble reproduced human preference at ρ ≈ 0.77 (image) and ρ ≈ 0.91 (group) without any guidance. Second, baseline agreement with paddy excluded (ρ = 0.772/0.911) was essentially equivalent to theory-directed agreement with paddy retained (ρ = 0.775/0.929), indicating that the theory-directed prompt improved AI–human alignment almost entirely by correcting the paddy undervaluation rather than by re-weighting all images.

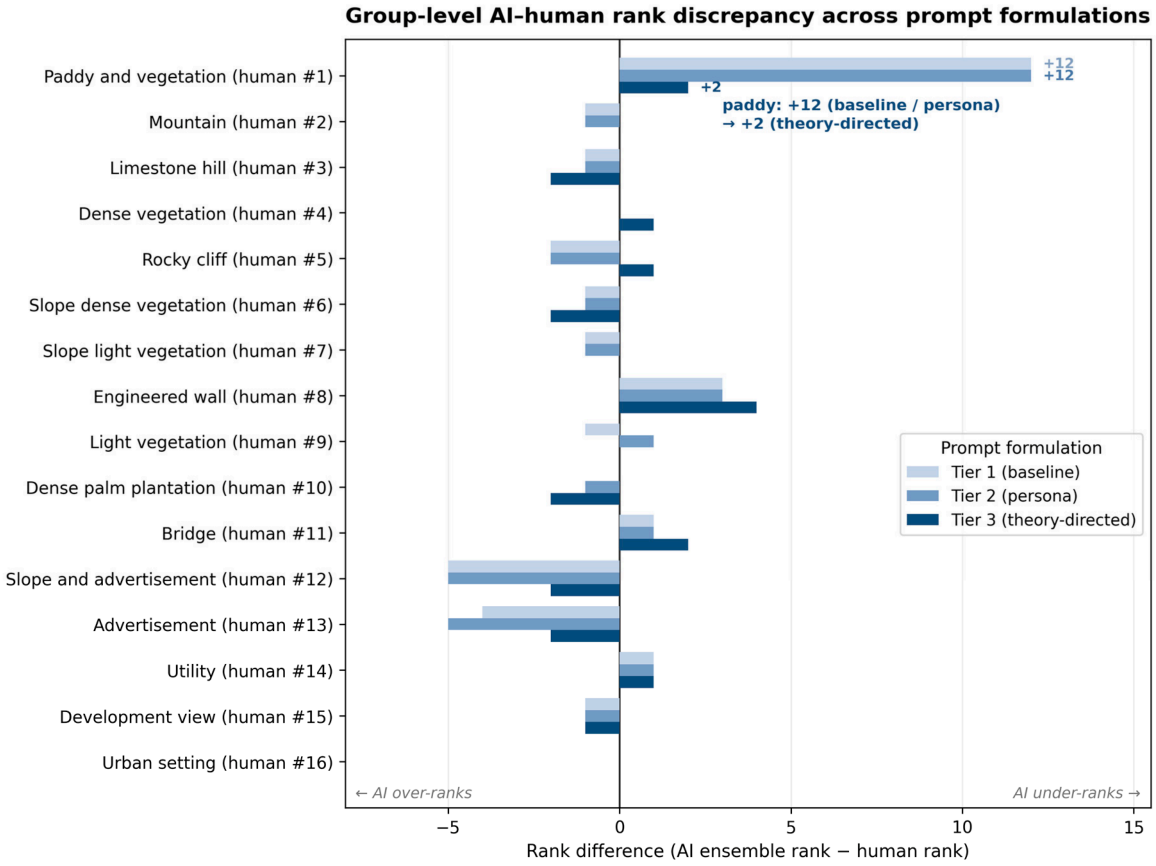


**Figure 7.** Group-level AI–human preference-rank discrepancy across the 16 landscape character groups under the three prompt formulations (Tier 1 baseline, Tier 2 persona, Tier 3 theory-directed).

### *3.7. Instrument Calibration*

This section tests H5, which predicted that AI–human agreement would be affected by whether human and AI preferences are elicited using the same response format. A potential confound in the foregoing comparison is that the AI rankings were elicited via forced choice, whereas the human benchmark (Section 3.2) was elicited via a rating scale. To separate genuine AI–human divergence from this difference in measurement, we collected an independent set of human forced-choice judgements on a 48-image validation subset: 60 respondents each completed 50 pairwise comparisons (3000 judgements in total), from which a human pairwise ranking was estimated using the same Bradley–Terry procedure applied to the models, with each choice weighted equally. All correlations in this section are computed over these 48 images (Table 6).

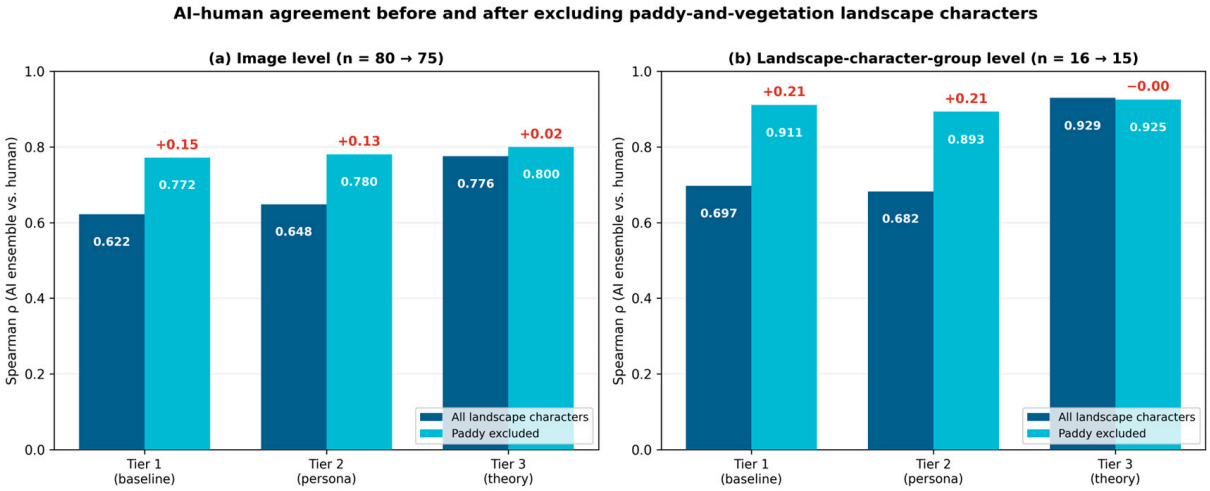


**Figure 8.** AI–human agreement (Spearman's ρ) before and after excluding the paddy-and-vegetation images, at (**a**) the image level and (**b**) the group level, for each prompt formulation. Red arrows mark the change.

**Table 6.** Instrument calibration on the 48-image validation subset: agreement between human rating-scale and forced-choice rankings, and between AI and human rankings within and across measurement paradigms.

| Comparison | Spearman's ρ [95% CI] | Kendall τ | Pearson r |
|---|---|---|---|
| **(a) Human–human, across instruments and samples** | | | |
| Likert ratings ($N$ = 400) vs. forced-choice pairwise ($N$ = 60) | 0.803 [0.64, 0.90] | 0.628 | 0.798 |
| **(b) AI ensemble vs. human, within the same (pairwise) paradigm** | | | |
| Tier 1 (baseline) | 0.842 [0.69, 0.93] | 0.688 | 0.849 |
| Tier 2 (persona) | 0.851 [0.71, 0.93] | 0.686 | 0.852 |
| Tier 3 (theory) | 0.840 [0.69, 0.92] | 0.663 | 0.861 |
| **(c) AI ensemble vs. human, across paradigms (pairwise AI vs. Likert human; cf. Section 3.6)** | | | |
| Tier 1 (baseline) | 0.630 | - | - |
| Tier 2 (persona) | 0.645 | - | - |
| Tier 3 (theory) | 0.774 | - | - |

Note. All correlations are computed over the 48 images evaluated in the pairwise validation study, at the individual image level ($n$ = 48) rather than aggregated to groups.

The two human instruments agreed strongly but not perfectly. The rating-scale ranking ($N$ = 400) and the forced-choice ranking ($N$ = 60) correlated at ρ = 0.803 (95% CI [0.64, 0.90]; Kendall's τ = 0.628, Pearson's r = 0.798; Figure 9a). Because these two rankings come from two independent respondent groups as well as two different elicitation methods, this value reflects both an instrument effect and a sampling effect; it is therefore best read not as pure measurement noise but as a realistic upper bound on the agreement attainable between two human landscape-preference rankings obtained under different conditions. In other words, even when two groups of human raters are measured in two ways, they place the 48 images in the same order, only to within ρ ≈ 0.80.

This reference reframes the AI–human comparison of Section 3.6. When the AI ensemble is compared with the human ranking elicited under the same (pairwise) paradigm, agreement is substantially higher than the cross-paradigm figures reported earlier: ρ = 0.842, 0.851, and 0.840 under the baseline, persona, and theory-directed prompts, respectively (Table 6b; Figure 9b), against the corresponding cross-paradigm values of ρ = 0.630, 0.645,

and 0.774 (Table 6c). Two implications follow. First, much of the apparent AI–human divergence observed in Section 3.6 is attributable to the mismatch in measurement instrument rather than to the models themselves: once humans are measured by forced choice, the ensemble already matches them at $\rho \approx 0.84$—at least as closely as the two human samples match each other ($\rho = 0.803$). Second, the advantage of theory-directed prompting seen in Section 3.6 does not persist within a common paradigm: the three prompt formulations agree with the human pairwise ranking almost identically ($\rho = 0.840$–$0.851$), indicating that the gain attributed to theory-directed prompting in the cross-paradigm comparison largely reflects its compensating for the rating-versus-choice mismatch, not a further improvement once that mismatch is removed. Taken together, these results indicate that the MLLM ensemble's landscape-preference ordering converges with human preference to a degree comparable to human-to-human agreement, and that the choice of elicitation instrument is a material factor when comparing AI and human assessments.

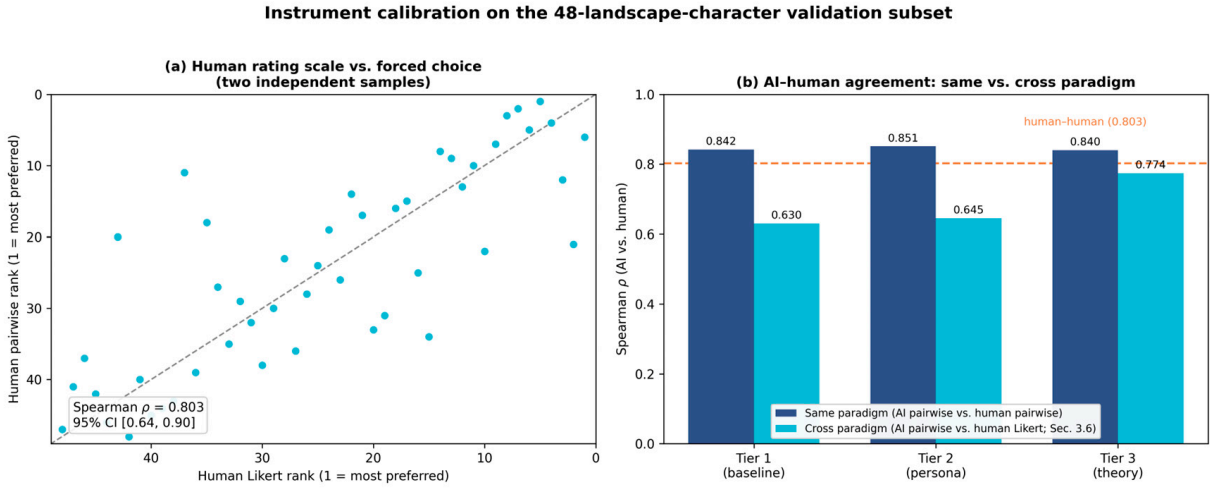


**Figure 9.** Instrument calibration on the 48 image validation subset. (**a**) Human rating-scale versus forced-choice rankings, and (**b**) AI–human agreement within versus across measurement paradigms.

## 4. Discussion

### *4.1. Principal Findings*

Five findings, each corresponding to one of the study's five hypotheses, emerged from this comparison.

H1 was supported. The five independently developed MLLMs converged on a highly concordant ranking of highway landscape preference (Kendall's W = 0.926 under the baseline prompt; pairwise $\rho = 0.86$–$0.95$), indicating that contemporary commercial MLLMs recover a stable ordinal structure of visual preference under controlled pairwise comparison, despite differences in training data, architecture, and position bias profiles.

H2 was supported. The AI ensemble's preference ranking was not fixed but varied with prompt formulation. The baseline and persona prompts produced rankings that were closely similar, whereas the theory-directed prompt, foregrounding openness and naturalness, substantially reordered the middle of the ranking; most notably, the paddy-and-vegetation group rose from rank 13 to rank 3 across all five models. The mechanism underlying this prompt sensitivity is examined in Section 4.3.

H3 was supported but qualified. The AI ensemble showed moderate agreement with human preferences under the baseline prompt at the image level ($\rho = 0.622$) and stronger agreement at the group level ($\rho = 0.697$). This agreement was itself sensitive to the prompt-induced reordering noted above: under the theory-directed prompt, image-level agreement rose to $\rho = 0.775$. The gap between AI–AI convergence (W = 0.926) and

AI–human convergence indicates that AI and humans share a broadly similar ordering but diverge within specific groups.

H4 was supported. AI–human divergences were systematic rather than random and concentrated in two opposing groups: paddy-and-vegetation images were strongly undervalued by AI relative to humans, while advertisement-related images were overvalued. Excluding the single paddy-and-vegetation group lifted image-level agreement from $\rho = 0.622$ to 0.772 (and group-level agreement from 0.697 to 0.911), confirming that a small subset of groups accounted for a disproportionate share of the total discrepancy.

H5 was supported. Measured AI–human agreement at the image level depended on the match between the human and AI rankings in the response format. When the human benchmark was elicited by forced choice—the same paradigm as the AI—AI–human agreement rose to $\rho \approx 0.84$, against $\rho \approx 0.63$ when the human benchmark was rating-based (Section 3.7), indicating that a substantial part of the apparent image-level AI–human divergence is an artefact of the asymmetric elicitation format rather than a genuine difference in preference.

Taken together, these findings cohere into a single picture. Contemporary commercial MLLMs recover a stable preference ordering (H1) that reproduces the broad shape of human highway-landscape preference; once the human benchmark is elicited under the same forced-choice paradigm, the ensemble agrees with human preference about as closely as two human samples agree with each other (H5). This alignment coexists with a specific, systematic blind spot—the undervaluation of cultivated paddy landscapes—that is neither random noise nor an intrinsic limitation, and is both diagnosable and correctable through prompt design (H2–H4). Current MLLMs are thus best understood as a credible, scalable complement to human landscape assessment rather than a substitute: reliable across most of the preference spectrum, but dependent on human judgement and deliberate prompting for the specific groups on which they diverge.

### *4.2. Convergent Machine Preference and Its Interpretation*

The high level of cross-model concordance indicates that contemporary MLLMs can recover a relatively stable ordinal structure of visual landscape preference under controlled comparison conditions. This should not be interpreted as evidence of human-like landscape appreciation, since surface-level behavioural alignment between large models and humans does not entail shared underlying mechanisms [40,41]. Rather, the convergence suggests that many highway landscape preferences are associated with visually explicit cues, such as topographic distinctiveness, vegetation density and layering, and the degree of urbanisation or infrastructure dominance, which are consistently encoded across large visual models.

This convergence extends prior work, which has largely inferred MLLM landscape or streetscape preference from a single model—for example, GPT-4 for urban attractiveness in Helsinki [20] and GPT-4o for streetscape aesthetics in Chongqing [25]—or, when deploying multiple models, has benchmarked each against human ratings independently rather than quantifying cross-model concordance [42]. The present concordance across five independent systems ($W = 0.926$) indicates that the broad preference ordering reported in those single-model studies is not model-specific and generalises to the current generation of commercial MLLMs.

The results also support the methodological value of pairwise comparison for MLLM-based visual preference assessment. Forced-choice comparison produces lower measurement variance and higher statistical sensitivity than single- and double-stimulus categorical scaling [33]; in the present study, it enabled complete preference ranking via Bradley–Terry modelling and, critically, allowed position bias to be diagnosed through AB/BA reversal—a safeguard unavailable in absolute rating designs.

Bias profiles nonetheless differed sharply across models. ChatGPT's residual flips were near-balanced, Claude was recency-biased, and Gemini, Kimi, and Qwen were primacy-biased under the baseline prompt; Qwen was the only model whose direction cleanly reversed across formulations, shifting to recency under the theory-directed prompt. The five models also differed in ways relevant to their use as landscape evaluators. ChatGPT was the most positionally reliable (consistent on 93.3% of pairs) and the most position-fair, making it the strongest single choice where only one model is affordable; Claude was the least consistent and the most idiosyncratic in its rankings, but this same distinctiveness makes it a useful ensemble member by diversifying error. Kimi and Qwen behaved most similarly to each other, so their agreement should not be mistaken for independent corroboration. No single model dominated on all criteria, which is the central practical argument for the majority-vote ensemble adopted here: it averages over the model-specific position biases and idiosyncrasies that, as Section 3.3 shows, none of the models is individually free of.

These visual-domain directions partially track text-domain findings: evaluating fifteen LLM judges on MTBench and DevBench, Shi et al. [24] reported GPT-4 as nearly preference-fair (PF = 0.02), Claude-3.5-Sonnet as recency-biased (PF = 0.22), and Qwen2.5-72B as near-fair (PF = −0.01), matching the directions observed here. Gemini, however, reversed across domains: Gemini-1.5-pro was recency-biased in the text domain (PF = 0.23 on MTBench) but primacy-biased in the present visual task, consistent with Shi et al.'s finding that preference direction is highly volatile across tasks (no text-domain benchmark is available for Kimi). This cross-model and cross-task variability is precisely why positional resolution mattered: without AB/BA reversal, these model-specific tendencies would have been indistinguishable from substantive landscape preferences, particularly for the strongly directional models.

Convergence, however, is not the same as correctness. Contemporary MLLMs are trained on heavily overlapping corpora and aligned through broadly similar procedures [19], and models trained on similar data tend to acquire correlated biases and errors—standardised across systems rather than diversified away [43]—a dynamic conceptually parallel to algorithmic monoculture, in which a shared evaluative signal can lower aggregate decision quality even when individually most accurate [44]. Cross-model consensus, therefore, indexes correlated structure in how these systems were built and aligned, not the independence that mutual corroboration would require.

This pattern also carries a methodological implication. The models' strong mutual agreement was obtained under a purely discriminative protocol—repeated "which is better?" judgements—rather than absolute score assignment. Recent work on LLM-based evaluation supports the value of this design: forced-choice pairwise assessment tends to align more closely with human judgement than independent pointwise scoring, particularly because the comparative framing makes fine distinctions easier to render [45,46], and, in the visual domain, forced-choice methods are more statistically efficient than rating scales [33]. The advantage is not unconditional, however: pairwise protocols are more susceptible to certain biases—including position effects and superficial stylistic distractors—and can exaggerate small differences relative to absolute scoring [47]. We therefore interpret our results as indicating that pairwise protocols offer practical advantages for MLLM-based landscape evaluation—chiefly position-bias diagnosis and cross-model comparability—rather than that MLLMs are inherently weaker at absolute scoring. A direct within-model comparison of the two paradigms and their respective bias profiles is a valuable direction for future work.

### *4.3. The Paddy Undervaluation as a Correctable Attentional Deficit*

The models' collective preference structure and the paddy undervaluation are two faces of the same mechanism. As the cross-model results show, the AI ensemble organises landscape preference primarily along vertical visual complexity—topographic relief, layering, and strong focal objects—rather than along horizontal openness, which human raters value. Paddy landscapes score low on the former and high on the latter, so their systematic undervaluation is a direct consequence of what the models attend to, not an isolated anomaly.

Malaysian road-landscape surveys report that open paddy views receive higher preference ratings than other rural landscape features [48], and Japanese rural-perception research finds that cultivated paddy landscapes are positively valued by both farming and non-farming stakeholders, with openness and peacefulness as significant predictors of preference [49]. The high value that Malaysian road users place on paddy landscapes is consistent with a pattern documented elsewhere in Southeast Asia, where rice cultivation is tied to cultural identity, collective memory, and heritage rather than to scenic drama alone. Across the region, paddy systems from the Balinese subak landscape to the terraces of the Philippine Cordillera are inscribed on the UNESCO World Heritage List as living cultural landscapes [50,51] and are valued for the openness, tranquillity, and sense of rural continuity they afford [15]; their significance is understood to rest on socio-ecological adaptation and ritual continuity as much as on visual form [52]. This cultural loading is precisely the kind of meaning that an image-only model, lacking situated regional experience, is least equipped to recover—an interpretive gap distinct from the perceptual one, and one that no amount of visual resolution alone would close.

The MLLMs, however, ranked the paddy group 13th, while the groups they most favoured—mountain, limestone hill, and rocky cliff—share the topographic drama, vertical mass, and strong contrast that paddy landscapes lack. The AI preference gradient thus tracks vertical complexity more strongly than horizontal openness: the two rankings agree at the extremes but diverge across the middle band, where the models resolve preference along a dimension human raters do not privilege. Figure 10 shows the individual scenes that most strongly drive this divergence: those that the AI ensemble over-ranked relative to road users and those it under-ranked.

Yet the same openness that human viewers prize is liable to be encoded by MLLMs as visual simplicity in the absence of dramatic topography or dense vegetation. Two converging accounts of how these models process characters explain this undervaluation. First, vision–language models encode object-level semantics more reliably than the spatial and relational structure that binds objects together, representing a scene as a loosely ordered collection of recognised content rather than as a configured whole [53]. A paddy character's appeal derives substantially from its openness, depth, and uninterrupted horizontal sweep—relational and configurational properties—rather than from any single salient object. Hence, a model anchored on object content has little to attach high value to. By the same logic, an advertisement-dominated character presents a large, legible, high-contrast object that a content-oriented reader can register as informative, even though human viewers experience roadside advertising as visual clutter that degrades scenic quality [54,55]; MLLMs may respond positively precisely because such signage increases local contrast, object diversity, and textural complexity. This resonates with deep-learning research showing that conventional deep models rely on local texture and surface-level statistical regularities rather than higher-order semantic structure when making visual judgements [56]. Although MLLMs differ architecturally from those convolutional networks, the tendency to privilege easily extractable signals over contextually grounded meaning may persist in multimodal systems that inherit vision encoders pretrained on similar objectives. Be-

yond these model-side accounts, a third explanation concerns what human raters bring to the task. Respondents evaluated the images against the lived experience of the corridor (Section 3.1), so their ratings plausibly encode place attachment and landscape memory alongside the visual stimulus [57], whereas MLLMs assess only the decontextualized image. This asymmetry is intrinsic to a road-user benchmark rather than a distortion of it.


**Figure 10.** Characters with the largest AI–human rank discrepancies (baseline prompt): scenes over-ranked by the AI ensemble (**top**) and under-ranked (**bottom**), each annotated with human and AI ranks (out of 80).

Second, the paddy characters were visually dominated by an expansive sky, flat terrain, and sparse foreground elements, yielding a low-complexity profile with limited tonal contrast and minimal vertical relief. These properties run counter to the dimensions that computational aesthetic models reward—compositional diversity, tonal contrast, and structural complexity [58]—and align with the midrange scores that large-scale aesthetic datasets assign to visually simple characters [18]. MLLMs, whose visual representations are shaped by web-scale corpora reflecting these tendencies, would therefore be expected to assign low preference scores to paddy characters.

The training corpora of all five models are proprietary and undisclosed. However, web-scale image datasets are commonly filtered using aesthetic predictors trained on photography-community ratings [18,59], in which calm, low-arousal themes rank among the lowest-scoring [18]—a convention that would underrepresent flat agricultural scenes among high-aesthetic training exemplars. Notably, the paddy undervaluation persisted across all five independently developed models, suggesting the bias reflects curation practices common to the current MLLM paradigm rather than the data provenance of any single corpus.

Critically, this deficit proved correctable rather than intrinsic. The theory-directed prompt, which foregrounded openness and naturalness as evaluation criteria, elevated the paddy group from rank 13 to 3 across all five models. Because the stimulus set was held fixed and only the evaluation criterion varied across the three prompt formulations, this reversal constitutes a controlled test of the attentional account. Persona framing alone (Tier 2) left the undervaluation essentially unchanged, whereas foregrounding openness and naturalness (Tier 3) reversed it, isolating what the models attend to rather than what they can perceive as the operative factor. As established in Section 3.6, this gain was essentially equivalent to that obtained by excluding the paddy group altogether, indicating

that the correction operated almost entirely through the paddy undervaluation rather than through a re-weighting of all characters. That so brief a redirection of attention toward configurational and naturalistic cues reversed a twelve-position error implies that the relevant information was available to the models all along but was not weighted as human viewers do. The divergence is thus better characterised as an attentional or framing deficit rather than an inability to perceive these qualities. Mechanistically, foregrounding openness and naturalness redirects the models' attention away from object content and compositional complexity toward configurational openness, so that the reversal follows from a shift in what is weighted rather than from any change in the scenes themselves. This accords with, and refines, reports that stakeholder- or theory-informed prompting substantially improves MLLM–human correlation [16] by isolating the specific landscape character and perceptual criterion through which the gain is realised.

### *4.4. Measurement Paradigm as a Source of Apparent Divergence*

A second source of apparent AI–human divergence was methodological rather than perceptual. A potential confound in the image-level comparison of Section 3.6 is that the AI ranking was elicited by forced choice while the human benchmark was elicited by a rating scale. When that comparison was re-estimated on the pairwise validation subset (Section 3.7), agreement between the ensemble and the human ranking rose from $\rho \approx 0.63$ under this cross-paradigm setup to $\rho \approx 0.84$ once the human benchmark was itself elicited by forced choice. This same-paradigm value is statistically indistinguishable from the agreement between two independent human samples measured by the two different instruments ($\rho = 0.803$), which itself sets a realistic ceiling on the agreement attainable between any two human landscape-preference rankings obtained under differing conditions. A material share of what would otherwise be read as AI–human divergence is therefore an artefact of comparing across elicitation formats rather than a genuine difference in preference.

Two implications follow. Methodologically, AI–human comparison studies that pit model forced choices against human ratings—or vice versa—risk overstating divergence, and the magnitude of that overstatement here ($\Delta\rho \approx 0.21$ at the image level) is comparable to the divergence the study set out to explain. Placing both judges on a common forced-choice footing, as the pairwise design does, removes a confound that absolute-rating studies cannot diagnose and simultaneously controls the position bias documented in Section 3.3 and in the broader LLM-as-a-judge literature [23,24]. Substantively, the finding tempers the apparent advantage of theory-directed prompting: within a common paradigm, all three prompt formulations agreed with the human pairwise ranking almost identically ($\rho = 0.840$–$0.851$), indicating that the cross-paradigm gain attributed to theory-directed prompting in Section 3.6 largely reflects its compensating for the rating-versus-choice mismatch rather than a further improvement once that mismatch is removed. Prompt engineering and instrument matching are thus partially substitutable routes to the same end, and reported gains from either should be interpreted in light of the elicitation format against which they are measured.

This nuance also qualifies the undervaluation of paddy in Section 4.3. Human raters place the paddy group first only on the rating scale; under forced choice, their placement of paddy shifts closer to the models' own placement, which is why the baseline ensemble already matches the human pairwise ranking at $\rho \approx 0.84$ without any corrective prompting. The undervaluation characterised in Section 4.3 is therefore most acute against rating-based benchmarks—the standard instrument in landscape-preference research—and attenuates once both judges are placed on a common forced-choice footing. Theory-directed prompting and instrument matching thus address overlapping portions of the same divergence rather than two independent ones.

### *4.5. Practical Implications: A Hybrid AI-Screening, Human-Targeted Workflow*

Taken together, the pattern of strengths and weaknesses documented here suggests a division of labour rather than a wholesale substitution of human assessment by MLLMs, and motivates a three-stage workflow for corridor-scale visual planning. The findings indicate where each judge is reliable: the ensemble agreed with human raters with high stability at the extremes of the distribution—consistently identifying geomorphologically distinctive landforms as most preferred and urban, utility, and development characters as least preferred, regardless of prompt wording (Section 3.5)—while its systematic errors were confined to a small, identifiable set of characters.

In the first stage, the MLLM ensemble screens the full image corpus along the corridor: pairwise evaluation is scalable to the tens of thousands of comparisons that corridor lengths demand but that human surveys cannot feasibly cover [16], and the ensemble's reliability at the extremes means clearly high- and clearly low-preference segments can be triaged automatically with confidence. Indeed, once the paddy group is set aside, the baseline ensemble reproduces human preference at $\rho \approx 0.77$ at the image level and $\rho \approx 0.91$ at the group level without any guidance (Section 3.6), so the extremes and the non-contentious middle of the corridor can be classified reliably before any human effort is spent. In the second stage, the characters known to be systematically misjudged—here, cultivated-agrarian landscapes such as paddy, which are undervalued, and advertisement-dominated characters, which are overvalued—are flagged for targeted human review (for the perceptual reasons set out in Section 4.3). This flag matters chiefly when the planning benchmark is a rating-based preference score, the standard instrument in landscape assessment: as Section 4.4 shows, the paddy undervaluation is most acute against rating scales and narrows when human input is itself collected by forced choice. A theory-directed prompt foregrounding openness and naturalness can be applied at the screening stage to reduce, though not eliminate, the paddy error against such rating-based benchmarks in advance. In the third stage, a focused human survey is administered only to the flagged and ambiguous mid-ranked segments, concentrating limited participant effort where machine judgement is least trustworthy rather than dispersing it uniformly across a corridor whose extremes the models already rank reliably. In operational terms, segments are flagged for human review when their ensemble rank departs from the human benchmark by more than a preset threshold (four to five group-level positions here) or when they belong to culturally loaded groups; because closed-weight models update silently, the screening stage should be re-calibrated against a small human anchor sample after each model version change. The result reserves human judgement for the culturally and configurationally loaded cases that current MLLMs do not, by default, evaluate as people do, while letting the ensemble carry the scalable remainder.

## 5. Limitations and Future Studies

Three categories of limitation bound the present conclusions and indicate directions for further work. Accordingly, the findings should be read as applying to still-image evaluation of a single tropical corridor by the current generation of commercial MLLMs, rather than to in-motion highway perception, to other landscape regions, or to MLLMs as a class.

The stimulus set was drawn from a single highway corridor in Peninsular Malaysia and presented as static Google Street View frames, which standardise capture conditions but omit the motion, peripheral flow, and temporal sequence that characterise the actual highway-viewing experience [5]. Whether the specific divergences observed here—particularly the paddy undervaluation—generalise to other corridors, regions, and agrarian landscape traditions remains to be tested, since paddy was among the strongest divergences in the dataset. Because real highway perception apprehends roadside elements

peripherally and sequentially rather than as composed frames [60], human preference under these conditions is arguably shaped less by fine composition than by rapidly integrated impressions of openness, enclosure, and rhythm [61]. We conjecture that the models' object-anchored bias—already evident in still images—would be amplified under dynamic input, since salient transient objects such as signage dominate attention when viewing time is compressed. In contrast, the gradual spatial qualities that humans integrate over a sequence would be harder for a frame-based model to recover. Testing MLLMs on video or image-sequence stimuli is, therefore, a natural next step. Concretely, the static protocol could be paired with short (5–10 s) forward-facing driving clips sampled at matched locations across three or more corridors; the same five models would then rate both the static and the video version of each site, and a static-versus-dynamic $\times$ landscape-group analysis would test directly whether the object-anchored bias is amplified under motion.

The human benchmark, while large ($N$ = 400), was recruited via snowball sampling. The forced-choice sample ($N$ = 60) that underpins H5 and the instrument-artefact interpretation (Sections 3.7 and 4.4) is a further limitation: it is smaller, independent of the main sample, and evaluated over only a 48-image subset. It establishes that the elicitation format materially affects measured AI–human agreement, but the precise magnitude of the same-paradigm gain should be interpreted as indicative rather than definitive until replicated in a larger matched sample. A direct replication would recruit a single respondent pool (target N $\geq$ 200) and randomly assign each participant to either the rating or the forced-choice format over the full 80-image set, yielding a within-study, between-subjects contrast that estimates the same-paradigm gain without the confound of two independently recruited samples.

Five commercial MLLMs were evaluated at a single point in time. Because such models are frequently updated and often opaque, the specific bias profiles reported here—including the positional-consistency ordering and the paddy undervaluation—may shift with future releases, and the findings are best read as a characterisation of the current model generation rather than of MLLMs in general. No open-source or task-fine-tuned models were included; fine-tuning on landscape-perception data offers a complementary route to the prompt-based correction examined here and warrants direct comparison in future work. Such a comparison could fine-tune one open-weight vision–language model on paired image–preference data (for example, the 400-respondent ratings collected here) and benchmark it against the same base model under the three-tier prompt on a held-out corridor, isolating whether parameter-level adaptation outperforms prompt-level redirection and whether it generalises beyond the training distribution. Reporting at the ensemble level deliberately averages over individual-model differences; while this stabilizes the preference estimate, it can mask model-specific behaviour of independent interest, as the divergent positional-bias profiles in Section 3.3 illustrate.

Three prompt formulations were compared, and the theory-directed prompt was constructed around openness and naturalness—criteria chosen post hoc to target the divergence identified here. This tailoring demonstrates that the paddy error is correctable in principle but does not establish that a single prompt will correct divergences of other kinds, in other landscape types, or in other cultural settings; there is a corresponding risk of overfitting the prompt to the particular error it was designed to address, and, because part of the measured benefit is specific to the rating-based benchmark against which it was assessed (Section 4.4), its value under other elicitation formats cannot be assumed. Whether theory-directed prompting confers a general alignment benefit or only a targeted correction of specific known biases is an open question that future work—systematically varying the theoretical framing and the landscape types to which it is applied—will need to resolve. A factorial design would cross several theoretical framings (e.g., prospect–refuge,

attention-restoration, and a neutral control) against multiple landscape domains (agrarian, urban, coastal, forest), measuring whether each framing improves alignment only for the domain it targets or transfers across domains, thereby distinguishing a general alignment benefit from targeted bias correction.

## 6. Conclusions

This study compared MLLM-derived and human visual preference rankings for 80 highway landscape images across 16 character groups along Malaysia's North–South Expressway. Five independently developed MLLMs converged on a highly concordant preference ranking (Kendall's W = 0.926) and reproduced the broad shape of human preference, agreeing moderately at the image level (Spearman's ρ = 0.622) and more strongly at the group level (ρ = 0.697) under the baseline prompt. The largest divergence—the paddy reversal, in which humans ranked cultivated paddy landscapes first while the models ranked them thirteenth—accounted for most of the AI–human gap; excluding this single group raised agreement to ρ = 0.772 and 0.911, respectively.

Two methodological findings qualify how this gap should be read. First, the complete pairwise design with AB/BA reversal proved essential: it enabled full Bradley–Terry ranking and showed that position bias is model-specific in both magnitude and direction, so that without reversal these tendencies would be indistinguishable from substantive preference. Second, the gap depended heavily on the elicitation format: when the human benchmark was itself collected via forced-choice rather than rating, agreement rose to ρ = 0.84, as close as two human samples measured by different instruments (ρ = 0.803). Much of the apparent divergence—and much of the benefit attributed to theory-directed prompting—therefore reflects a rating-versus-choice mismatch rather than a genuine perceptual difference. Taken together, these findings reframe what AI–human "divergence" in landscape preference actually consists of: once the measurement paradigm is matched, alignment is close, and what remains is not diffuse unreliability but a narrow, identifiable, and prompt-correctable blind spot. MLLMs reliably track visually explicit cues—topographic distinctiveness, vegetation density, infrastructure dominance—but, against the rating-based benchmarks standard in landscape research, undervalue landscapes whose appeal rests on openness and cultural familiarity and overvalue advertisement-dominated scenes that humans read as scenic contamination. This alignment is therefore feature- and instrument-conditional rather than uniform—a profile that positions MLLMs as a scalable screening layer for corridor-wide visual inventory rather than a substitute for human preference survey, with human judgement reserved for the culturally loaded landscape character groups, and the rating-based comparisons, on which the models are known to diverge. Whether this conditional alignment narrows as models advance is a question the present design is equipped to track.

**Supplementary Materials:** The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/land15071240/s1, Table S1: Representative images of the 16 highway landscape character groups; Table S2: Demographic and highway-use profile of the 400 survey respondents; Table S3: Mean visual preference scores for all 80 individual highway landscape images ($N$ = 400); Table S4: Positional-bias (primacy/recency) breakdown for the five MLLMs across the three prompt formulations; Table S5: Pairwise cross-model Kendall's τ values across the three prompt formulations; Table S6: Intransitivity (cyclic-triple) rates for the five MLLMs across the three prompt formulations; Table S7: Five-model ensemble Bradley–Terry scores for all 80 images across the three prompt formulations.

**Author Contributions:** Conceptualization, H.G., R.S. and S.A.B.; methodology, H.G., R.S. and S.M.; software, H.G. and J.Y.; validation, H.G. and S.M.; formal analysis, H.G. and J.Y.; investigation, H.G.; resources, S.A.B.; data curation, H.G.; writing—original draft preparation, H.G.; writing—review and editing, H.G., R.S., S.A.B., S.M. and J.Y.; visualization, H.G. and J.Y.; supervision, R.S., S.A.B. and S.M. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in the study.

**Data Availability Statement:** The MLLM pairwise judgement dataset supporting this study is openly available on Figshare at https://doi.org/10.6084/m9.figshare.32332245.

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

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| AI | Artificial intelligence |
| API | Application programming interface |
| AB/BA | Original (AB) and reversed (BA) presentation orders of an image pair |
| BT | Bradley–Terry (model) |
| CI | Confidence interval |
| MLLM | Multimodal large language model |
| NSE | North–South Expressway |

## Appendix A

| Prompt Tier | Prompt Text (Verbatim) |
|---|---|
| Tier 1 (Baseline) | You are comparing two highway landscape scenes along Malaysia's North–South Expressway. Which scene do you find more visually preferable? Focus only on scenic quality. Ignore weather, traffic, and road engineering. Answer with a single letter only, A or B, and nothing else. |
| Tier 2 (Persona) | You are a road user who has travelled along the northern section of Malaysia's North–South Expressway at least once in the past year, whether driving or riding as a passenger. You are comparing two highway landscape scenes along Malaysia's North–South Expressway. Which scene do you find more visually preferable? Focus only on scenic quality. Ignore weather, traffic, and road engineering. Answer with a single letter only, A or B, and nothing else. |
| Tier 3 (Theory-directed) | You are a road user who has travelled along the northern section of Malaysia's North–South Expressway at least once in the past year, whether driving or riding as a passenger. You are comparing two highway landscape scenes along Malaysia's North–South Expressway. Which scene do you find more visually preferable? When judging, consider two aspects that often shape how appealing a scene looks. First, scenes are generally favoured when the view is open, and the eye can survey into the distance, while still offering some sense of enclosure or shelter. Second, scenes are generally valued for their naturalness, and conspicuous human-made elements that fail to blend into the scene tend to lower their perceived scenic quality. Weigh how open and how natural each scene is in this light, and respond as you genuinely see it. Focus only on scenic quality. Ignore weather, traffic, and road engineering. Answer with a single letter only, A or B, and nothing else. |

## References


1. Appleyard, D.; Lynch, K.; Myer, J.R. *The View from the Road*; MIT Press: Cambridge, MA, USA, 1964.
2. Chamberlain, B.C.; Meitner, M.J. A route-based visibility analysis for landscape management. *Landsc. Urban Plan.* **2013**, *111*, 13–24. [CrossRef]
3. Kent, R.L. Determining scenic quality along highways: A cognitive approach. *Landsc. Urban Plan.* **1993**, *27*, 29–45. [CrossRef]
4. Martín, B.; Arce, R.; Otero, I.; Loro, M. Visual landscape quality as viewed from motorways in Spain. *Sustainability* **2018**, *10*, 2592. [CrossRef]
5. Topolšek, D.; Areh, I.; Cvahte, T. Examination of driver detection of roadside traffic signs and advertisements using eye tracking. *Transp. Res. Part F Traffic Psychol. Behav.* **2016**, *43*, 212–224. [CrossRef]
6. Qin, X.; Fang, M.; Yang, D.; Wangari, V.W. Quantitative evaluation of attraction intensity of highway landscape visual elements based on dynamic perception. *Environ. Impact Assess. Rev.* **2023**, *100*, 107081. [CrossRef]
7. Zhu, M.; Xu, J.; Jiang, N.; Li, J.; Fan, Y. Impacts of road corridors on urban landscape pattern: A gradient analysis with changing grain size in Shanghai, China. *Landsc. Ecol.* **2006**, *21*, 723–734. [CrossRef]
8. Forman, R.T.T. Town ecology: For the land of towns and villages. *Landsc. Ecol.* **2019**, *34*, 2209–2211. [CrossRef]
9. Li, C.; Zhang, J.; Philbin, S.P.; Yang, X.; Dong, Z.; Hong, J.; Ballesteros-Pérez, P. Evaluating the impact of highway construction projects on landscape ecological risks in high altitude plateaus. *Sci. Rep.* **2022**, *12*, 5170. [CrossRef] [PubMed]
10. Zube, E.H.; Sell, J.L.; Taylor, J.G. Landscape perception: Research, application and theory. *Landsc. Plan.* **1982**, *9*, 1–33. [CrossRef]
11. Daniel, T.C. Whither scenic beauty? Visual landscape quality assessment in the 21st century. *Landsc. Urban Plan.* **2001**, *54*, 267–281. [CrossRef]
12. Dramstad, W.; Tveit, M.S.; Fjellstad, W.; Fry, G. Relationships between visual landscape preferences and map-based indicators of landscape structure. *Landsc. Urban Plan.* **2006**, *78*, 465–474. [CrossRef]
13. Jaal, Z.; Abdullah, J. User's preferences of highway landscapes in Malaysia: A review and analysis of the literature. *Procedia Soc. Behav. Sci.* **2012**, *36*, 265–272. [CrossRef]
14. Jaal, Z.; Abdullah, J.; Ismail, H. Malaysian North South Expressway landscape character: Analysis of users' preference of highway landscape elements. *WIT Trans. Ecol. Environ.* **2013**, *179*, 365–376. [CrossRef]
15. Tekken, V.; Spangenberg, J.H.; Burkhard, B.; Escalada, M.; Stoll-Kleemann, S.; Truong, D.T.; Settele, J. "Things are different now": Farmer perceptions of cultural ecosystem services of traditional rice landscapes in Vietnam and the Philippines. *Ecosyst. Serv.* **2017**, *25*, 153–166. [CrossRef]
16. Lin, J.; Chen, C.; Feng, T.; Huo, S.; Zhang, K.; Dong, B.; Xiang, S.; Wang, K.; Huang, L. Exploring the potential of generative AI to complement multi-stakeholder landscape preference assessment. *Landsc. Urban Plan.* **2026**, *270*, 105602. [CrossRef]
17. Tung, Y.; Yang, Z.; Shen, M.; Chang, C.; Chen, C.; Ho, L. Research note: Assessing human preferences for natural landscapes—An analysis of ChatGPT-4 and LLaVA models. *Landsc. Urban Plan.* **2025**, *259*, 105371. [CrossRef]
18. Murray, N.; Marchesotti, L.; Perronnin, F. AVA: A large-scale database for aesthetic visual analysis. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–21 June 2012*; IEEE: Piscataway, NJ, USA, 2012; pp. 2408–2415. [CrossRef]
19. Yin, S.; Fu, C.; Zhao, S.; Li, K.; Sun, X.; Xu, T.; Chen, E. A survey on multimodal large language models. *Natl. Sci. Rev.* **2024**, *11*, nwae403. [CrossRef] [PubMed]
20. Malekzadeh, M.; Willberg, E.; Torkko, J.; Toivonen, T. Urban attractiveness according to ChatGPT: Contrasting AI and human insights. *Comput. Environ. Urban Syst.* **2025**, *117*, 102243. [CrossRef]
21. Liu, J.; Zhang, T.; Ma, Y.; Hu, T.; Lin, F.; Liang, H.; Yang, D.; Pan, Y.; Gao, D.; Qiu, L.; et al. Generative artificial intelligence perspectives on typical landscape types: Can ChatGPT compete with human insight? *Landsc. Urban Plan.* **2025**, *264*, 105479. [CrossRef]
22. Li, H.; Zhang, Y.; Ding, J.; Li, Q.; Zhang, P. Visual Room 2.0: Seeing is not understanding for MLLMs. *arXiv* **2025**, arXiv:2511.12928.
23. Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.P.; et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 46595–46623. [CrossRef]
24. Shi, L.; Ma, C.; Liang, W.; Diao, X.; Ma, W.; Vosoughi, S. Judging the judges: A systematic study of position bias in LLM-as-a-judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL 2025)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 292–314.
25. Zhou, Q.; Zhang, J.; Zhu, Z. Evaluating urban visual attractiveness perception using multimodal large language model and street view images. *Buildings* **2025**, *15*, 2970. [CrossRef]
26. Yoo, M.; Cho, G.-H.; Kim, D. Explaining Walkability with a Zero-Shot Multimodal LLM: Validation and Empirical Application. *Sustain. Cities Soc.* **2026**, *144*, 107431. [CrossRef]
27. Zhang, J.; Li, Y.; Fukuda, T.; Wang, B. Urban Safety Perception Assessments via Integrating Multimodal Large Language Models with Street View Images. *Cities* **2025**, *165*, 106122. [CrossRef]

28. Ma, H.; Kwan, M.-P. Measuring the Psychological Restorative Quality of Urban Spaces: A Vision Language Model-Based Method. *Sci. Rep.* **2026**, *16*, 16534. [CrossRef] [PubMed]
29. Wicke, P.; Wachowiak, L. Exploring spatial schema intuitions in large language and vision models. *arXiv* **2024**, arXiv:2402.00956. [CrossRef]
30. Yao, Y.; Dall'Ò, G.; Lu, F. Urban Street-Scene Perception and Renewal Strategies Powered by Vision–Language Models. *Land.* **2026**, *15*, 244. [CrossRef]
31. Anonymous. Details omitted for double-blind peer review (companion classification study). *J. Asian Archit. Build. Eng.* 2026, *manuscript under review*.
32. Kaplan, R.; Kaplan, S. *The Experience of Nature: A Psychological Perspective*; Cambridge University Press: Cambridge, UK, 1989.
33. Mantiuk, R.K.; Tomaszewska, A.; Mantiuk, R. Comparison of four subjective methods for image quality assessment. *Comput. Graph. Forum* **2012**, *31*, 2478–2491. [CrossRef]
34. Bradley, R.A.; Terry, M.E. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika* **1952**, *39*, 324–345. [CrossRef]
35. He, H.; Li, S.; Liu, J.; Li, S.; Shen, H.T. Analyzing and boosting the power of fine-grained visual recognition for multimodal large language models. *arXiv* **2025**, arXiv:2501.15140.
36. Appleton, J. *The Experience of Landscape*; Wiley: London, UK, 1975.
37. Kaplan, S. The restorative benefits of nature: Toward an integrative framework. *J. Environ. Psychol.* **1995**, *15*, 169–182. [CrossRef]
38. Gao, H.; Bakar, S.A.; Maulan, S.; Yusof, M.J.M.; Mundher, R.; Guo, Y.; Chen, B. A systematic literature review and analysis of visual pollution. *Land* **2024**, *13*, 994. [CrossRef]
39. Maystre, L.; Grossglauser, M. Fast and accurate inference of Plackett–Luce models. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 172–180.
40. Mitchell, M.; Krakauer, D.C. The debate over understanding in AI's large language models. *Proc. Natl. Acad. Sci. USA* **2023**, *120*, e2215907120. [CrossRef] [PubMed]
41. Strachan, J.W.A.; Albergo, D.; Borghini, G.; Pansardi, O.; Scaliti, E.; Gupta, S.; Saxena, K.; Rufo, A.; Panzeri, S.; Manzi, G.; et al. Testing theory of mind in large language models and humans. *Nat. Hum. Behav.* **2024**, *8*, 1285–1295. [CrossRef] [PubMed]
42. Belaroussi, R.; Martín-Gutiérrez, J. Architectural ambiance: ChatGPT versus human perception. *Electronics* **2025**, *14*, 2184. [CrossRef]
43. Bommasani, R.; Hudson, D.A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M.S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. On the opportunities and risks of foundation models. *arXiv* **2021**, arXiv:2108.07258.
44. Kleinberg, J.; Raghavan, M. Algorithmic monoculture and social welfare. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2018340118. [CrossRef] [PubMed]
45. Liusie, A.; Manakul, P.; Gales, M.J.F. LLM Comparative Assessment: Zero-Shot NLG Evaluation Through Pairwise Comparisons Using Large Language Models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), St. Julian's, Malta, 17–22 March 2024*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 139–151.
46. Liu, Y.; Zhou, H.; Guo, Z.; Shareghi, E.; Vulić, I.; Korhonen, A.; Collier, N. Aligning with Human Judgement: The Role of Pairwise Preference in Large Language Model Evaluators. In Proceedings of the First Conference on Language Modeling (COLM 2024), Philadelphia, PA, USA, 7–9 October 2024.
47. Tripathi, T.; Wadhwa, M.; Durrett, G.; Niekum, S. Pairwise or Pointwise? Evaluating Feedback Protocols for Bias in LLM-Based Evaluation. In Proceedings of the Second Conference on Language Modeling (COLM 2025), Montreal, QC, Canada, 7–10 October 2025.
48. Gao, H.; Bakar, S.A.; Maulan, S.; Yusof, M.J.M.; Mundher, R.; Zakariya, K. Identifying visual quality of rural road landscape character by using public preference and heatmap analysis in Sabak Bernam, Malaysia. *Land* **2023**, *12*, 1440. [CrossRef]
49. Natori, Y.; Chenoweth, R. Differences in rural landscape perceptions and preferences between farmers and naturalists. *J. Environ. Psychol.* **2008**, *28*, 250–267. [CrossRef]
50. UNESCO. Rice Terraces of the Philippine Cordilleras (World Heritage List No. 722). Available online: https://whc.unesco.org/en/list/722 (accessed on 1 July 2026).
51. UNESCO. Cultural Landscape of Bali Province: The Subak System as a Manifestation of the Tri Hita Karana Philosophy (World Heritage List No. 1194). Available online: https://whc.unesco.org/en/list/1194 (accessed on 1 July 2026).
52. Acabado, S.; Albano, A.; Martin, M. Conservation for Whom? Archaeology, Heritage Policy, and Livelihoods in the Ifugao Rice Terraces. *Land.* **2025**, *14*, 1721. [CrossRef]
53. Kamath, A.; Hessel, J.; Chang, K.-W. What's "up" with vision-language models? Investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), Singapore, 6–10 December 2023*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 9161–9175. [CrossRef]
54. Portella, A. *Visual Pollution: Advertising, Signage and Environmental Quality*; Routledge: London, UK, 2016.

55. Wakil, K.; Tahir, A.; Hussnain, M.Q.u.; Waheed, A.; Nawaz, R. Mitigating urban visual pollution through a multistakeholder spatial decision support system to optimize locational potential of billboards. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 60. [CrossRef]
56. Geirhos, R.; Jacobsen, J.-H.; Michaelis, C.; Zemel, R.; Brendel, W.; Bethge, M.; Wichmann, F.A. Shortcut learning in deep neural networks. *Nat. Mach. Intell.* **2020**, *2*, 665–673. [CrossRef]
57. Scannell, L.; Gifford, R. Defining place attachment: A tripartite organizing framework. *J. Environ. Psychol.* **2010**, *30*, 1–10. [CrossRef]
58. Deng, Y.; Loy, C.C.; Tang, X. Image aesthetic assessment: An experimental survey. *IEEE Signal Process. Mag.* **2017**, *34*, 80–106. [CrossRef]
59. Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. LAION-5B: An open large-scale dataset for training next generation image-text models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 25278–25294. [CrossRef]
60. Yang, S.; Wilson, K.; Roady, T.; Kuo, J.; Lenné, M.G. Beyond Gaze Fixation: Modeling Peripheral Vision in Relation to Speed, Tesla Autopilot, Cognitive Load, and Age in Highway Driving. *Accid. Anal. Prev.* **2022**, *171*, 106670. [CrossRef] [PubMed]
61. Blumentrath, C.; Tveit, M.S. Visual Characteristics of Roads: A Literature Review of People's Perception and Norwegian Design Practice. *Transp. Res. Part A Policy Pract.* **2014**, *59*, 58–71. [CrossRef]