Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Ranking-Based Human Validation Layer for Concept Implementation in AI-Generative Design

Manuel Ibáñez-Arnal

manuel.ibanez@uchceu.es

Universidad Cardenal Herrera CEU

Luis Doménech-Ballester

Universidad Cardenal Herrera CEU

Victor García-Peñas

Universidad Cardenal Herrera CEU

Article

Keywords:

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11627821/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Ranking-Based Human Validation Layer for Concept Implementation in AI-Generative Design.

Manuel Ibáñez-Arnal*, Luis Doménech-Ballester and Victor García-Peñas

Department of Mathematics, Physics and Technological Sciences, University CEU Cardenal Herrera, CEU Universities, Avda. Seminario s/n, 46113 Moncada, Valencia, Spain

manuel.ibanez@uchceu.es *corresponding author

Abstract

Text-conditioned generative AI can produce graded series of product-form variants intended to express a target concept with increasing intensity, but whether human observers perceive such series as coherent gradients has rarely been tested. We propose a ranking-based protocol to test it. For nine Product–Concept pairs (e.g., futuristic sneaker, polygonal lamp), an image-to-image latent-diffusion workflow produced six variants intended to increase concept expression (A–F). Twenty-six experienced design engineers ranked each series by perceived concept intensity and could flag variants no longer recognisable as the product category (not applicable, NA). Rankings were analysed with inter-observer agreement, rank correlation with the intended order and Plackett–Luce aggregation with uncertainty, and NA frequencies were used to locate recognisability limits. Five pairs showed strong agreement (Kendall's $W \geq 0.89$) and recovered the intended order, two showed moderate agreement with local ambiguity between adjacent levels, and two showed no reliable ordering. In most pairs, NA responses increased towards the upper levels, with high agreement on which variants had lost category identity. These patterns define three perceptual regimes—stable, stable with weak segments and unstable—that indicate whether, and over which range, an AI-generated concept series is read consistently by expert observers before it is used in design exploration.

Introduction

Engineering design is an intentional activity in which product form serves not only functional requirements but also communicates meanings, values, and conceptual attributes that shape interpretation in use. During early design stages, designers frequently explore directions that can be operationalised as concept axes, such as aggressivity, elegance, or futurism, long before detailed performance optimisation becomes feasible. The usefulness of such conceptual directions depends on whether observers reliably perceive graded differences in the intended attribute across design variants. Traditional perceptual-evaluation approaches—such as semantic differentials and Kansei-based methods—have long enabled researchers to characterise perceived attributes, but they typically assess impressions of fixed designs rather than the interpretability of controlled conceptual variation. In this work, each target concept is treated operationally as a single conceptual axis for early-stage exploration; this is a practical representation for testing perceptual monotonicity rather than a theoretical claim about the concept's internal semantic structure.

The recent emergence of generative AI has dramatically expanded designers' ability to create large families of visually coherent alternatives at minimal cost. Yet increasing generative capacity introduces a new challenge: the ability to produce variation does not guarantee that this variation is perceptually coherent or monotonic. A sequence intended to progressively intensify a concept may be read inconsistently, non-monotonically, or in ways that diverge from the designer's semantic intent. For instance, a sneaker design intended to appear progressively "more futuristic" may do so initially, but further intensification can

distort proportions or silhouette cues to the point where observers no longer recognise the item as belonging to the same product category. When this happens, the series no longer supports meaningful comparison within the product category. Validating whether concept-driven manipulations produce a monotonic, interpretable gradient is therefore a precondition for using such a series as a basis for design decisions.

Until the emergence of generative AI models, intensifying a conceptual attribute in a product form—making it "more futuristic," "more silent," "more ruby-like"—was an exclusively human operation, mediated by the designer's interpretive judgement and manual craft. Contemporary text-conditioned generative AI introduces a practically new capacity for concept-driven design exploration: it can automatically produce families of product-form variants intended to intensify a target concept without requiring designers to first translate that concept into explicit geometric rules, CAD parameters, Kansei objective functions, or trained latent-space directions. This capacity raises an empirical question that precedes any claim of design control: whether the resulting variation is perceived by human observers as a monotonic, interpretable and recognisable gradient of the intended concept. The question is no longer only whether AI can generate alternatives, but whether a generated concept series is read consistently enough to inform early-stage design decisions.

Despite the practical relevance of this question, there is no lightweight, general method for determining whether AI-generated conceptual gradients are perceptually coherent, where ambiguity emerges, or where intensification undermines recognisability. Existing perceptual-evaluation approaches were not designed to diagnose the structure of parametric conceptual variation produced by generative models, and current generative-design research rarely tests whether conceptual manipulations behave predictably for human observers.

This work addresses that gap by proposing a ranking-based human validation layer for concept-driven generative workflows. The approach provides an efficient perceptual check

that determines whether a set of AI-generated variants intended to span increasing levels of a target concept is interpreted in the intended direction and with sufficient consistency to support early-stage design decisions. Rather than assessing preference or aesthetic appeal, the method focuses on interpretive directionality and consistency—whether observers perceive a coherent increase (or decrease) along the conceptual axis. In addition to recovering the intended ordering, we examine whether adjacent intensity levels are perceptually separable, identifying segments where local indistinguishability or inversions emerge. To handle extreme cases where conceptual intensification compromises category identity, a simple "not applicable" (NA) option marks recognisability failure, providing a signal of where intensification compromises category identity.

We evaluate the protocol in an online study with expert design engineers using multiple Product–Concept pairs. Each pair consists of a product category combined with a target conceptual attribute (e.g., "futuristic sneaker," "elegant lamp"), instantiated through a fixed generative workflow (Fig. 1) that produces a fixed number of intensity levels intended to represent a monotonic increase in concept expression. Rankings are processed to address the following research questions:

- RQ1 (Reliability): Do observers exhibit consistent orderings for a given Product–Concept pair?
- RQ2 (Validity): Does the aggregated human ordering align monotonically with the intended A–F progression?
- RQ3 (Operational range): Do NA responses emerge systematically—e.g., at higher intended levels—indicating bounds within which concept intensification remains recognisable and practically usable?

This paper makes three contributions to engineering design research and practice:

1. Methodological contribution. We introduce a lightweight, reproducible perceptual-validation protocol based on ranking with NA and ordinal–probabilistic analysis. This method enables design teams to evaluate whether a concept-intensification series generated by AI is perceived as a monotonic, interpretable and separable gradient—a question not directly addressed by existing product-semantics, Kansei or generative-AI evaluation frameworks.

2. Conceptual contribution. We propose a general diagnostic framework describing three perceptual regimes—R1 stable, R2 stable-with-weak-segments, and R3 unstable—capturing distinct ways in which AI-generated conceptual gradients manifest perceptually. This framework provides a shared vocabulary for describing how consistently a generated concept series is read by observers within the tested configuration.

3. Operational contribution. We reconceptualise loss of recognisability (NA) not as noise or missing data but as design-relevant evidence marking the boundary beyond which intensification compromises category identity. This perspective enables the explicit identification of the range of levels within which a generated series remains recognisable.

Background

Product semantics and conceptual intent in product form

Research in product semantics has long established that product form communicates meaning beyond technical function, positioning products as sign systems through which users infer category, affordances, values, and identity cues. Meaning emerges from the interpretive relation between observer and artefact rather than from form alone, implying that semantic

intent must be inferred through perception, interpretation, and use rather than directly observed[1]. Within this tradition, concept implementation can be understood as the intentional encoding of semantic cues in form so that a target concept becomes legible to observers. Yet, as Kazmierczak[2] argues, design operates as meaning-making embedded in culture, which makes the gap between designer intention and user interpretation inevitable. Cross-cultural research further reinforces this variability, showing that the same formal configuration can produce divergent meanings depending on social context.

An important implication for engineering design is that concept implementation is rarely binary. Designers often manipulate continuous degrees of an attribute—such as "more aerodynamic" or "more futuristic"—suggesting an underlying perceptual continuum rather than categorical shifts. The challenge is to vary semantic cues in a graded and predictable way without compromising functional legibility or recognisability, especially during early design stages when exploration is abundant but interpretive evidence remains limited.

Measuring perceived attributes: semantic differential, Kansei, and affective evaluation

Efforts to measure subjective meaning in design have evolved through several complementary approaches. The semantic differential, introduced by Osgood, Suci and Tannenbaum[3], provided the first systematic tool for mapping impressions of artefacts along bipolar adjective pairs, allowing researchers to quantify perceived connotations in a multidimensional "semantic space." Building on this foundation, Schütte[4] formalised Kansei Engineering, which seeks to translate affective responses into actionable design parameters, effectively linking user emotion to form features and enabling iterative refinement. These methods have been adapted to a wide range of product categories and cross-cultural contexts.

More recent developments integrate computational and data-driven methods to enhance semantic analysis. Siddharth, Blessing and Luo[5] demonstrated how natural language

processing can extract affective meaning from textual corpora, enriching the traditional semantic differential framework with algorithmic scalability. Similarly, Feng et al.[6] proposed an AI-assisted approach that integrates Kansei principles with knowledge-graph reasoning to generate user-centric interface designs, showing how semantic intent can guide generative prompts. Affective evaluation has also expanded toward physiological and behavioural dimensions; Guo et al.[7] combined psychophysiological signals with perceptual ratings to strengthen empirical grounding in emotional design studies. Yet despite these advances, there remains a persistent need for lightweight, scalable methods capable of capturing meaningful perceptual distinctions in early-stage concept evaluation—without requiring extensive instrumentation.

Ranking and comparative judgement in perceptual evaluation

When constructs are inherently subjective or absolute calibration is unreliable, comparative methods such as paired comparisons, sorting, and ranking provide a more natural means of eliciting perceptual information. Ranking-based evaluation aligns closely with the way designers reason about form—making relative judgements such as "A expresses the concept more than B" rather than assigning arbitrary numeric scores. Empirical evidence suggests that relative methods can reduce scale-use bias and increase sensitivity in perceptual data[8,9]. From a modelling perspective, probabilistic frameworks such as the Plackett–Luce or Bradley–Terry models[10–12] estimate latent "worth" parameters underlying ranked preferences. These models are interpretable, robust to incomplete data, and well-suited for aggregating ordinal information across observers. Importantly, they also handle the common challenge of recognisability breakdown—when extreme design manipulations distort category cues to the point that items become incomparable. In concept-driven design gradients, this property allows analysts to treat perceptual collapse not as error but as a meaningful limit of interpretability.

Generative AI in conceptual design and the need for validation

The rise of generative AI has revolutionised conceptual design by enabling the creation of large sets of visually coherent form alternatives at minimal cost. These systems have been shown to augment designers' creativity, broaden the search space, and accelerate iteration[13,14]. However, this proliferation of generative capacity introduces a new bottleneck: validation. The essential question is shifting from "Can AI generate alternatives?" to "Do these alternatives reliably implement the intended conceptual change in a way that humans interpret as intended?"

Empirical evidence underscores this challenge. In the AutoSpark study, Chen et al.[13] integrated Kansei Engineering principles into a generative AI pipeline for automotive appearance design, demonstrating that affective intent can be encoded and evaluated systematically. Liang[15] extended this line of inquiry to cultural and creative products, proposing adaptive validation frameworks sensitive to audience diversity. Luo[16] developed a multi-modal evaluation system combining visual, textual, and affective data to verify whether AI-generated forms convey the desired concept with perceptual coherence. Collectively, these works show that generative novelty alone is insufficient for engineering design—what matters is whether conceptual expression changes in an interpretable and monotonic way across levels of manipulation.

Hybrid AI–human validation and human-in-the-loop perspectives

Hybrid AI–human validation frameworks position human judgement not as auxiliary data collection but as an integral part of generative workflows. Human–AI collaboration research consistently emphasises that humans contribute contextual knowledge, interpretive calibration, and correction mechanisms that align model outputs with human goals[17]. In design contexts, Fattah Saleh[18] illustrated how industrial designers integrate subjective reasoning and affective assessment within generative pipelines, while Wang et al.[19] proposed

a co-annotation framework where human feedback iteratively refines AI-generated aesthetic variations. Theoretically, Comte[20] argues that this reframing repositions designers as semantic mediators rather than autonomous creators—agents who ensure that AI systems remain accountable to human interpretive logic. Consequently, validation becomes the procedural bridge that transforms algorithmic exploration into evidence-based design knowledge, ensuring that generated artefacts remain both perceptually legible and semantically meaningful.

Positioning of the present work

Building on the traditions of product semantics, Kansei Engineering, and affective evaluation, the present work proposes a hybrid AI–human validation layer for concept-driven generative workflows. The method is designed to empirically test whether generative manipulations produce perceptually monotonic gradients of concept expression while maintaining recognisability. Unlike aesthetic preference studies, which often assess appeal or liking, the focus here is on interpretive consistency—whether a given concept (e.g., "aggressive," "elegant") is perceived as increasing or decreasing in strength across generated levels. Using ranking-based perceptual evaluation analysed through probabilistic models, the framework enables ordinal aggregation, uncertainty quantification, and cross-concept comparability. Moreover, by explicitly encoding "not applicable" or "unrecognisable" responses, the approach treats recognisability breakdown as an informative boundary condition—marking where concept expression begins to compromise category identity.

This perceptual monotonicity is crucial: if observers do not consistently perceive an increase (or decrease) in the intended concept across levels, then the generated series cannot be relied upon to represent graded differences in that concept, and design comparisons based on it become misleading.

Methods

Study overview and experimental design

This study evaluates whether AI-generated conceptual gradients are interpreted by human observers in the intended direction. The experimental unit is a Product–Concept pair, for which a controlled generative workflow produced six visual variants (A–F) representing an ordinal increasing intended expression of a target concept, without assuming perceptual equidistance between adjacent levels. Participants completed a perceptual ranking task, providing an ordered sequence from lowest to highest perceived concept expression.

Six levels were selected as a practical compromise between perceptual resolution and cognitive load. Fewer levels would have provided insufficient granularity to detect local inversions or recognisability breakdown, whereas adding more levels increases the combinatorial complexity of the ranking task and the likelihood of perceptual ties. The A–F scale therefore supports the methodological goals of detecting monotonicity, ambiguity, and recognisability loss without overloading participants.

An explicit recognisability mechanism was included to identify points at which conceptual intensification undermines product-category identity. For any stimulus that was no longer perceived as a valid instance of the target product category, participants were instructed to assign the label Not Applicable (NA). NA is strictly defined as a categorical judgement of recognisability loss and must not be used for any other reason—such as uncertainty, difficulty discriminating between adjacent items, or perceived ties. As Fig.1 shows, when a stimulus was marked NA, it was removed from the ordering interface, and participants ranked the remaining items. This mechanism enables the study to characterise the range of levels within which the generated series remains recognisable as the target product.

Stimuli generation

Stimuli were AI-generated product-form images produced under a controlled workflow designed to modulate a target concept across six intended levels (A–F). For each Product–Concept pair, the product category was held constant while the conceptual attribute was parametrically intensified across levels using a consistent base prompt structure and fixed visual constraints (viewpoint, framing, background, and lighting) to support comparability. This yields a monotonic progression along the generative-pipeline axis, with A–F corresponding to successively intensified concept prompts from the system's parametrization.

The generative pipeline was implemented in ComfyUI using a latent-diffusion image-to-image architecture (Figure 2). An input image is VAE-encoded into an initial latent representation; in parallel, positive and negative text prompts are embedded via the CLIP text encoder to provide semantic conditioning and constraint signals. Using an SDXL model checkpoint (UNet + CLIP + VAE), the diffusion sampler (KSampler) performs iterative denoising in latent space (20 steps, CFG = 8, Euler sampler, normal scheduler, fixed seed; and a denoise sweep) to generate a refined latent, which is then VAE-decoded back to image space and saved as the final output. No post-processing was applied beyond standard file export (PNG encoding) by ComfyUI. SDXL was selected as the generative backbone because it is an open-source, locally executable model. This supports long-term reproducibility, as the workflow can be re-run without dependence on commercial APIs.

The selection of these specific KSampler settings was informed by prior exploratory studies conducted by the authors, in which the principal parameters of the image-to-image pipeline—including CFG scale, step count, sampler type, scheduler, seed strategy, and denoise strength—were systematically varied and their effects on concept attribution assessed. These preliminary calibrations identified which parameters exerted the most significant influence on the perceived intensity and coherence of conceptual attributes in the generated forms, and

guided the configuration adopted in the present study. A detailed account of this parametric exploration is beyond the scope of this paper and will be reported separately.

For each level, several candidate images were generated under identical sampling settings, and one representative image per level was selected by the authors using criteria defined before the perceptual study: (i) product-category fidelity, (ii) absence of unintended artefacts, (iii) compliance with the fixed visual constraints and (iv) consistency with the intended level of concept intensity. Criteria (i)–(iii) concern image quality and comparability. Criterion (iv) entails a judgement of concept intensity by the authors and is therefore a potential source of experimenter bias; the perceptual study is designed to test this judgement rather than presuppose it. Had curation imposed an ordering that observers did not share, this would surface as low agreement or misalignment with the intended order—as observed for two of the nine pairs. The implications of this curation step are discussed under Limitations. The complete ComfyUI workflow, the input base images, the prompt specifications for each level and all stimuli used in the study are openly available (see Data availability).

The study included nine Product–Concept pairs: Bottle–Bubble, Bottle–Origami, Vase–Ruby, Lamp–Polygonal, Table–Contrast, Chair–Aerodynamic, Chair–Aerodynamic 2, Sneaker–Silence, and Sneakers–Futuristic. These pairs were not intended as a statistically representative sample of the product–concept universe—an aim that no finite set of cases could meaningfully fulfil—but as a set of heterogeneous theoretical probes designed to exercise the validation logic across diverse conditions of conceptual design practice. The input images serving as the base for the image-to-image pipeline were real product photographs selected by the researchers according to deliberate design criteria intended to minimise perceptual confounds. Base images depicted products against neutral backgrounds, exhibiting pure and elementary formal attributes with no salient pre-existing conceptual connotations that could bias observers' subsequent rankings—except in cases where such

connotations were deliberately retained for research purposes. Products were required to be highly recognisable exemplars drawn from a range of product categories. Purely functionalist products were explicitly excluded on the grounds that strong form–function coupling could lead raters to conflate usability cues with aesthetic-conceptual perception, thereby undermining the validity of the ranking task. The selected products were therefore those in which styling plays a prominent role while still departing from minimal, archetypal forms. Base images were selected by the authors according to these criteria before any stimuli were generated. The criteria concern the suitability of each starting product as a carrier of styling variation, not the expected outcome of the ranking task; no independent rating of the base images was conducted. The pairs span multiple product categories (bottles, vases, lamps, tables, chairs, sneakers) and concept types (geometric, stylistic, metaphorical, perceptual), enabling the validation approach to be exercised across a wide range of scenarios representative of conceptual design practice. Across pairs, the intent was to test the same validation logic under heterogeneous conditions typical of conceptual design practice, rather than to optimise performance for a single object or concept or to support statistical generalisation across the product–concept space.

The generative pipeline may modify multiple formal attributes simultaneously when intensifying a conceptual prompt, given the entangled nature of latent generative representations. We do not aim to isolate or manipulate a single semantic feature in a controlled factorial sense. Instead, our interest lies in assessing whether the aggregate effect of the prompt intensification results in a perceptually coherent directional gradient as interpreted by human observers. In this framework, any auxiliary or emergent variations introduced by the model—such as changes in curvature, sharpness, colour accents or textural cues—are considered part of the model's expressive mechanism for rendering the specified concept and are therefore accepted as natural within the evaluation protocol.

Survey instrument and procedure

The evaluation was implemented through an online drag-and-drop interface. For each trial, the six stimuli of a Product–Concept pair were displayed simultaneously in a randomised layout. Participants were instructed to order the items from lowest to highest perceived expression of the target concept, forming a strict total order without ties. If two variants appeared similar, participants were asked to select the ordering that best reflected even subtle differences. Figure 3 shows the response collection platform.

In accordance with the definition provided previously, participants could mark any stimulus as NA when it was not recognisable as belonging to the target product category. NA items were automatically removed from the interface, and the participant completed a partial ranking with the remaining stimuli.

Participants confirmed their ordering before proceeding to the next trial. All instructions were provided in Spanish, matching participants' native language.

Participants and data collection

This study received ethics approval from the Ethics Committee of the Universidad Cardenal Herrera-CEU (CEEI 25/705). All procedures were performed in accordance with the relevant guidelines and regulations, and informed consent was obtained from all participants before they started the survey.

Participants were qualified design engineers (≥5 years' experience), recruited to ensure familiarity with product-form cues and reduce ambiguity in concept-expression judgments. The final sample comprised N = 26 qualified design engineers. Participant professional experience in engineering design ranged from 5 to 22 years (*M = 12.23, SD = 4.04, Mdn = 13.5*). Recruitment was conducted via professional networks, alumni networks and industry contacts, and all participants provided informed consent prior to starting the survey. Responses were collected remotely through the online survey, completed on participants' own

devices. Participation was voluntary and responses were anonymised prior to analysis. The study was conducted in Spanish with native Spanish-speaking participants; all task instructions and concept labels were presented in Spanish.

The linguistic and cultural homogeneity of the participant sample was a deliberate methodological choice. By ensuring that both the researchers who designed and administered the study and the observers who evaluated the stimuli shared the same native language and cultural background, any potential perceptual or interpretive gap arising from cross-cultural or cross-linguistic differences was controlled for at the design stage. This decision prioritises internal validity: since conceptual labels such as "futuristic," "silence," or "origami" may carry different connotations across languages and cultural contexts, a shared linguistic and cultural frame between emitters and receivers ensures that the rankings reflect differences in concept-form perception rather than differences in semantic interpretation of the task labels. The implications of this choice for external generalisability are acknowledged in the Limitations section.

Twenty-six qualified design engineers participated in the study. Sample sizes of this magnitude are common in expert-panel research in design semantics, Kansei Engineering, and perceptual evaluation, where recruitment is constrained by the need for domain-specific expertise[4,21,22]. Because each complete ranking implicitly encodes a dense set of pairwise comparisons, ranking-based tasks provide high informational yield per participant, reducing the need for large samples compared to rating-scale methodologies[8]. Comparative judgements have further been shown to yield more reliable subjective image assessments than rating scales[9]. Moreover, the aim of the present study is methodological validation—evaluating the perceptual structure of concept-intensification gradients—rather than population-level estimation. For this type of objective, expert samples in the range of 20–30 participants are typical and sufficient in prior design-research literature.

Data structure and preprocessing

For each participant and Product–Concept pair, the response is an ordered list of level labels drawn from the set {A, B, C, D, E, F}, assigned to ordinal positions 1–6. NA-labelled items create missing positions, producing partial rankings whose length varies across participants and pairs. Raw survey outputs were validated to ensure internal consistency (e.g., no duplicate assignments, no impossible position allocations), then converted into an ordinal representation suitable for rank-based statistics and probabilistic modelling.

Rankings were treated as directional according to the task instruction (lower→higher). NA responses were retained as an explicit outcome variable rather than discarded. This separation supports later analyses that distinguish disagreements about concept expression from failures of category recognition.

(1) Visual consensus diagnostics: For each Product–Concept pair, a level-by-position contingency matrix was computed, counting how often each nominal level (A–F) was placed in each ranking position (1–6). These matrices were visualised as heatmaps. Concentration along a diagonal pattern indicates convergence toward the intended monotonic ordering, while off-diagonal structure may reflect confusion, inversion, or local indistinguishability between adjacent levels.

Analysis of NA responses and operational range

In line with RQ3, NA responses were treated as indicators of category-recognition breakdown and potential saturation in the concept-intensification process. For each Product–Concept pair, the proportion of NA responses was computed per nominal level (A–F). Rising NA frequencies at higher levels may be consistent with over-extension, whereas NA at lower levels would signal an early breakdown or floor-effect in which minimal manipulation

undermines category identity. This behaviour defines the operational range within which the conceptual axis remains recognisable and practically usable.

To evaluate whether recognisability breakdown occurs consistently across observers, we computed several agreement coefficients on the binary omission variable (NA vs. Non-NA) for each level. The observed agreement (Po) quantifies the raw proportion of coincident omission decisions. Fleiss' κ and Krippendorff's α estimate interobserver agreement beyond chance, with α being robust to the presence of missing values. Because NA can be highly prevalent or extremely rare depending on the Product–Concept pair, we also computed Gwet's AC1, which corrects for prevalence induced biases that typically deflate κ. High agreement indices indicate that observers converge on the same subset of levels as non-recognisable, supporting a well-defined operational boundary. Conversely, low agreement suggests heterogeneous omission criteria or the presence of subgroups operating in different recognisability regimes, in which case no shared operational boundary can be assumed.

In addition to agreement metrics, we evaluated whether omission patterns were systematically related to the intended level. First, a chi-square test of independence (NA vs. Non-NA × A–F) was computed to assess whether NA occurrence was randomly distributed across levels or whether certain levels concentrated omissions disproportionately. Second, to test for monotonic patterns in omission rates along the A→F progression, both the Cochran–Armitage trend test and Jonckheere-Terpstra test were computed using a two-sided alternative. Directionality (increasing vs. decreasing) was determined from the sign of the standardized Z statistic. A significant trend indicates that NA frequencies vary systematically across levels, revealing where recognisability tends to fail more frequently along the intended axis. Together with the agreement indices, these tests distinguish between (i) homogeneous, collectively shared boundaries of recognisability and (ii) heterogeneous or level-unspecific omission patterns.

Probabilistic aggregation and uncertainty estimation

To aggregate rankings and estimate uncertainty, a Plackett–Luce (PL) model was fitted for each Product–Concept pair. The PL model estimates a "worth" parameter for each level, reflecting its probability of being ranked above other levels under a probabilistic choice interpretation. This provides an interpretable aggregate ordering together with model-based estimates of uncertainty for each level. Because PL models naturally accommodate partial rankings, NA-induced omissions do not disrupt the estimation process.

Uncertainty estimation is particularly important for engineering design interpretation because adjacent levels may be perceptually close even when the overall ordering is monotonic. To formalise this idea, we refer to local separability (or adjacent-level separability) as the degree to which consecutive intended levels are perceptually differentiated. Whereas global monotonicity concerns the preservation of the overall A→F direction, local separability is a local property of the gradient that captures the potential overlap or ambiguity between neighbouring levels. Within the present framework, local separability is assessed through the uncertainty associated with the Plackett–Luce worth estimates and through the consecutive pairwise comparison probabilities derived from the model.

All analyses were implemented in a reproducible computational workflow. The manuscript reports, for each Product–Concept pair: (i) descriptive response distributions and heatmaps, (ii) agreement metrics, (iii) alignment statistics, (iv) PL worth estimates with uncertainty, and (v) NA patterns across levels. Given N = 26 observers, each pair yields up to 26 rankings, with reduced effective comparisons in cases where NA produces partial rankings.

Regime classification across the nine Product–Concept pairs was performed by the research team as a qualitative diagnostic judgement based on the joint reading of multiple complementary indicators rather than by applying numerical thresholds to any single statistic.

For each pair, the assignment considered: (i) the level of inter-observer agreement (Kendall's W, Krippendorff's α, Fleiss' κ); (ii) the alignment of the aggregated ordering with the intended A–F progression (Spearman's ρ, Kendall's τ); (iii) the local separability between adjacent levels as reflected in the Plackett–Luce adjacent-comparison probabilities and confidence intervals; and (iv) the visual structure of the level–position heatmaps. The classification reflects our methodological position that the perceptual structure of concept-driven generative gradients cannot be reduced to a single coefficient, and that diagnostic regimes are better characterised through the convergence of multiple indicators than through fixed numerical cut-offs. To support transparency and replicability, Table 1 reports the empirical profile of each regime as observed in the present dataset.

The indicator ranges reported in Table 1 are descriptive of the present dataset and should not be interpreted as universal cut-offs for regime assignment. Applying this framework to new product–concept pairs requires examining the same set of indicators jointly, and may yield empirical profiles that differ from those observed here as a function of concept type, product category, generative pipeline, and observer population. Table 1 is therefore intended as a transparent reference for the diagnostic judgement exercised in the present study, not as a substitute for it.

Outcome measures

The outcome analysis in this section addresses RQ1 and RQ2 by focusing on two complementary properties of the ranking data: interobserver reliability and directional validity. Reliability captures whether participants converge on a consistent ordering of the stimuli, whereas validity assesses whether that consensus—when present—follows the intended A–F progression defined by the generative pipeline. Distinguishing these two dimensions is essential, since participants may agree on an ordering that does not match the intended direction, or conversely, they may align with the direction while exhibiting low

overall agreement. This separation clarifies the respective roles of reliability as a precondition for interpretability and validity as evidence of perceptual monotonicity.

Reliability: inter-observer agreement. To quantify whether observers share a consistent ordering, agreement among participants was assessed using ordinal-appropriate inter-rater measures (e.g., Kendall's W; Krippendorff's alpha for ordinal data).

Validity: alignment with intended level order. To test whether human judgements align with the intended progression defined by the generative pipeline, rank-based association measures were computed between the aggregated human ordering and the A–F reference order (e.g., Spearman's ρ and/or Kendall's τ). High positive association indicates that participants perceive increasing concept expression in the intended direction; low or negative association may reflect, among other factors, weak interpretability, ambiguity, or systematic inversion of the intended concept axis. Because the aggregated order comprises only six levels, ρ and τ computed on it are coarse indicators; they are therefore interpreted jointly with inter-observer agreement and Plackett–Luce separability rather than in isolation, so that a perfect aggregate correlation with low agreement is not read as evidence of a reliable gradient.

The perceptual constructs examined in this study—agreement among observers, directional alignment with the intended A–F order, monotonicity, and separability between adjacent levels—are conceptually distinct and cannot be fully characterised by any single ordinal statistic. Different indicators capture different aspects of these constructs and rely on different assumptions regarding data structure, scale interpretation, and the presence of partial rankings. For this reason, the analysis incorporates a set of complementary measures rather than relying on a single index. This approach allows each construct to be assessed using tools

specifically suited to its properties, without implying that all indicators measure the same phenomenon or that they should lead to identical outcomes.

Data availability

The ComfyUI workflow, base images, prompt specifications, generated stimuli, anonymised ranking data and analysis scripts supporting this study are openly available in Zenodo at https://doi.org/10.5281/zenodo.18908998.

Use of large language models

Claude (Anthropic) was used to assist with language editing and the drafting of selected passages. All content was reviewed, revised and approved by the authors, who take full responsibility for it.

Results

Figure 4 displays the level-position heatmaps for all Product-Concept pairs. Each matrix shows the frequency with which levels A-F were placed in positions 1-6 across participants, with darker cells indicating higher counts. The diagonal shading patterns represent how often the intended ordering was recovered, while off-diagonal placements show deviations or local reassignments. The NA row captures items that were excluded from rankings due to loss of recognizability.

Figure 5 presents the estimated probability (from the ordinal model) that each higher-numbered level was placed above the preceding one (e.g., 2 over 1, 3 over 2). Bars near 1.0 indicate frequent consecutive ordering in the expected direction, whereas values closer to 0.5 reflect reduced local separability between those adjacent levels.

Figure 6 shows the log-worth estimates obtained from the Plackett-Luce aggregation for levels A-F, together with their confidence intervals. The relative horizontal position of the

estimates reflects the ordering recovered by the model, while the width and degree of interval overlap indicate the uncertainty associated with each level. Substantial overlap denotes limited separation between levels, whereas non-overlapping intervals correspond to clearer distinctions in the aggregated data.

Figure 7 reports the number of NA responses per participant and Product-Concept pair. Each cell indicates how many items an individual observer judged as not recognisable within the category, and row/column totals summarise overall omission patterns. The distribution of NA responses across participants helps determine whether recognisability issues were concentrated among specific individuals or shared more broadly.

Figure 8 displays the number of NA responses per intended level across all participants. For each Product-Concept pair, the bars show how often items at levels A-F were excluded from the ranking due to lack of recognisability. Differences in NA frequency across levels reveal whether omissions cluster at particular points of the intended gradient.

Table 2 summarises the omission (NA) patterns observed across the nine Product–Concept pairs. Overall, the proportion of coincident NA decisions (Po) was high, indicating that observers largely agreed on which stimuli were no longer recognisable as instances of the target category. Agreement was maximal in Table–Contrast and Chair–Aerodynamic, where Po = 1.00 and no variability in NA responses was recorded. Consequently, Fleiss' κ, χ^2, and other variation-dependent statistics could not be computed for these pairs.

For Bottle–Bubble, Bottle–Origami, Vase–Ruby, Lamp–Polygonal, and Sneakers–Futuristic, agreement coefficients were also high, with Po values ranging from 0.93 to 0.98 and Gwet's AC1 estimates between 0.86 and 0.95. In these pairs, the χ^2 tests produced very low p-values (≈ 0.0002), and both the Cochran–Armitage and Jonckheere–Terpstra trend tests yielded similarly significant results, indicating that NA responses were not evenly distributed across

levels A–F and instead showed clear level-dependent structure in their frequencies. The positive sign of the corresponding trend statistics reflects that NA frequencies increased across the intended A→F gradient—that is, higher levels exhibited more omissions, whereas lower levels exhibited fewer.

In contrast, Chair–Aerodynamic 2 and Sneaker–Silence exhibited low and scattered omission rates. This pattern is reflected in κ and α values close to zero or slightly negative, which arise from the minimal variability in NA decisions rather than from disagreement. Correspondingly, χ^2 and trend-test p-values were high (≥ 0.16), indicating that NA responses in these pairs were more uniformly distributed across levels and did not display evidence of a systematic trend.

Table 3 reports the reliability and validity metrics for the perceptual rankings across the nine Product–Concept pairs. Several pairs showed strong and coherent agreement among observers, with high Kendall's W, Krippendorff's α, and Fleiss' κ values—particularly Bottle–Origami, Vase–Ruby, Lamp–Polygonal, Table–Contrast, and Sneakers–Futuristic ($W \geq 0.89$). These same pairs also exhibited perfect or near-perfect rank-based correlations with the intended A–F ordering (Spearman $\rho = 1.00$; Kendall $\tau = 1.00$), indicating that participants consistently recovered the target monotonic progression.

Moderate levels of agreement were observed for Chair–Aerodynamic and Chair–Aerodynamic 2 ($0.59 \leq W \leq 0.72$), with Chair–Aerodynamic showing only partial alignment with the intended sequence ($\rho = 0.94$; $\tau = 0.87$). In these cases, rankings were generally directionally correct but displayed local variability in the relative placement of adjacent levels.

For Bottle–Bubble and Sneaker–Silence, inter-rater agreement was too low to support a meaningful validity assessment. Both pairs showed very weak reliability, with Kendall's $W \leq$

0.25, Krippendorff's $\alpha \leq 0.27$, and Fleiss' $\kappa \leq 0.17$. In these conditions, the dispersion and inconsistency of individual rankings prevent the interpretation of alignment metrics with the intended A–F order, and no conclusions regarding directional validity can be drawn for these pairs. Although the aggregated order for Bottle–Bubble numerically followed the intended A–F sequence ($\rho = 1.00$), the very low inter-observer agreement prevents this alignment from being interpreted as evidence of a reliable perceptual gradient.

Across all Product–Concept pairs, the patterns observed in Tables 2 and 3 were consistent with the visual evidence provided by the level–position heatmaps and by the Plackett–Luce log-worth confidence intervals. Pairs with high agreement and clear directional validity showed tight, predominantly diagonal structures in the heatmaps and non-overlapping log-worth intervals across levels. Conversely, the pairs with low agreement (Bottle–Bubble and Sneaker–Silence) displayed highly dispersed heatmaps and extensive overlap in log-worth intervals, mirroring the absence of a stable perceptual structure in their rankings.

Discussion

From statistical outputs to design-relevant regimes

The core question for engineering design is not whether a particular estimator (e.g., rank aggregation) behaves well, but whether a generated concept series is read predictably by human observers. Across our nine Product–Concept pairs we observe three recurrent regimes of perceptual behaviour for AI-generated gradients:

- R1-Stable: The intended A→F progression is recovered with strong consensus and clear adjacent-level separation. Pairs such as Lamp-Polygonal, Table-Contrast or Sneakers-Futuristic exemplify this regime.

- R2-Stable with weak segments: the global direction A→F holds, but one or two transitions show local ambiguity or partial inversions. Cases such as Chair-Aerodynamic or Chair-Aerodynamic 2 illustrate this behaviour.
- R3-Unstable: no reliable ordering emerges; placements disperse and the intended axis is not read consistently. This pattern appears in Bottle-Bubble or Sneaker-Silence

These regimes are qualitative patterns identified from the joint reading of agreement (W, α), directional alignment (ρ, τ), local separability (Plackett–Luce intervals and pairwise probabilities) and recognisability (NA patterns), rather than threshold-based classes (see Methods and Table 1). Users of the protocol may adopt stricter numerical criteria where their application requires them.

Orthogonal to this, NA behaviour marks category recognisability and therefore the operational range of the gradient: NA absent or late (E–F) indicates a broad usable span; NA early (C–D) indicates that intensification quickly undermines category identity. The R-regime answers "is the gradient interpretable and how uniformly?"; the NA pattern answers "how far can intensification proceed before the design ceases to read as the same product?" Together, R1–R2–R3 × NA provides a diagnostic frame for judging whether, and over which range, a generated concept series is read consistently enough to support early-stage design exploration.

These regimes emerge from comparing the monotonic generative axis with its perceptual counterpart. They do not demonstrate perceptual monotonicity; rather, they reveal the different ways in which a monotonic AI-driven manipulation can be interpreted—sometimes coherently, sometimes ambiguously, and sometimes not at all. We stress that these regimes are empirical regularities, not universal laws: they summarise how the combined system (pipeline + concept definition + product category) expresses a putative conceptual dimension

as read by humans. They are therefore testable across additional concepts, categories, and generative controls.

The perceptual regimes observed in this study (R1, R2, R3 × NA) arise from the specific generative pipeline employed—an SDXL-based latent diffusion model implemented in ComfyUI, with fixed prompts, sampler configuration and seed. As with any generative workflow, alternative models, checkpoints, schedulers or seed strategies may produce gradients with different levels of monotonicity, separability or recognisability. This model-dependence is inherent to current generative-AI systems, and therefore the regimes identified here should be interpreted as characteristic of this pipeline rather than as universal properties of concept-driven image synthesis. Nonetheless, a distinction should be drawn between the regime-specific findings reported here, which are properties of the tested SDXL-based configuration, and the diagnostic structure of the framework itself—capturing stability, ambiguity and breakdown through the four perceptual conditions identified in Section 5.5—which is not tied to a specific model and can be applied to other architectures, although this remains to be tested.

Sources of regime behaviour

The three perceptual regimes identified in this study—R1 stable, R2 stable-with-weak-segments, and R3 unstable—should be understood as empirical patterns emerging from the joint configuration of four interacting elements: the target concept, the product category, the generative pipeline (including model, prompts, sampler settings and seed strategy), and the observer population. None of these elements alone determines the resulting regime; regime behaviour is a property of the combined system rather than of any of its components considered in isolation. The factors outlined below offer plausible, non-exhaustive interpretations that help situate the behaviours observed in our dataset, without committing to a single causal locus for any specific case.

- R1 (stable gradients) tend to arise when the target concept has shared semantic structure within the observer population and the generative pipeline encodes a coherent latent direction for that concept and the product category provides sufficient structural tolerance for variation without compromising recognisability. Under these jointly favourable conditions, intensification reinforces perceptually aligned and category-compatible cues, producing smooth and discriminable transitions.
- R2 (stable with weak segments) tends to occur when the global concept-pipeline-category alignment is preserved but at least one factor introduces local irregularity. This may happen, for instance, when the concept is interpretable but internally multi-cue, leading the pipeline to intensify different attributes at different points along the gradient; when the pipeline produces non-uniform spacing between adjacent intensity levels; or when sensitive regions within the product category amplify small formal changes perceptually. The result is a globally coherent gradient with locally ambiguous or weakly separable segments.
- R3 (unstable gradients) tends to occur when one or more of the four elements fails to support reliable perceptual ordering under the conditions of the experiment. This may reflect a concept that lacks a shared perceptual interpretation in the observer population, a pipeline that does not encode a coherent latent direction for that concept under the prompts and parameters used, a product category whose structural constraints interact poorly with the intended manipulation, or a combination of these factors. R3 is therefore not evidence of an intrinsic incompatibility between concept and product; it is evidence that the tested concept–category–pipeline–observer configuration did not yield a reliable perceptual gradient. Different model architectures, prompt formulations, or observer populations may produce different

regimes for the same nominal product–concept pair, and distinguishing among these potential sources requires additional control conditions identified as future work.

Together, these factors clarify that regime behaviour is a diagnostic consequence of the combined system—concept, category, pipeline and observer—rather than a property of the validation method itself or of any single component. The framework's role is to detect and characterise this behaviour, not to attribute it.

Relation to prior work in Product Semantics and Kansei

The present findings align with long-standing perspectives in product semantics and Kansei Engineering, where product form communicates meaning through perceptual cues and cultural conventions. Traditional approaches—such as the semantic differential[3] and Kansei Engineering[4,21]—provide structured ways to measure affective impressions of fixed designs using rating scales. While these methods are highly effective for mapping connotative meaning, they do not assess whether systematic intensification along a single conceptual axis produces a perceptually coherent, monotonic gradient. Rating-based frameworks characterise impressions of individual artefacts but do not diagnose the interpretability, local separability, or breakdown of concept-driven variation.

Our contribution complements this tradition by shifting the analytical focus from evaluating impressions of discrete designs to assessing whether a generative manipulation produces a consistently perceived gradient. By incorporating local ambiguity and recognisability loss (NA) explicitly—phenomena that rating scales cannot capture—the proposed validation layer extends product-semantics and Kansei methodologies to graded, AI-driven manipulation of conceptual expression.

Whereas this situates our contribution within design-specific approaches to meaning and affect, the behaviour of the gradients also reflects more general perceptual regularities, which connects the present work to classical psychophysics.

Conceptual link to psychophysics: gradients, local discriminability, and breakdown

From a broader perceptual-science perspective, the gradient behaviours observed here parallel well-established properties of psychometric functions. Parametric manipulations typically yield monotonic but locally variable discriminability, with some transitions producing steep, easily perceptible differences and others yielding shallow or ambiguous perceptual changes[23]. Under certain conditions—such as noisy cues or competing attributes—local non-monotonicities may also emerge[24]. This structure maps directly onto our regimes: R1 resembles a well-formed psychometric function with consistently high discriminability, R2 corresponds to locally shallow or weakly separable segments, and R3 reflects situations in which no coherent perceptual ordering arises from the manipulation.

Similarly, the emergence of NA responses mirrors the collapse of category recognition observed at extreme stimulus values in psychophysical tasks, when distortions or cue degradation prevent observers from maintaining category identity[23]. Although we do not model psychometric functions explicitly, this analogy provides a principled foundation for interpreting stability, ambiguity, and breakdown in AI-generated conceptual gradients. It clarifies why certain transitions fail, why some segments remain ambiguous, and why intensification eventually undermines recognisability—insights relevant to how generated concept series are used in design exploration.

Actionable guidance for design teams

The validation layer functions as a diagnostic check of whether a generated concept series is read consistently within a given generative configuration. The combined reading of perceptual regime (R1–R2–R3) and recognisability range (NA pattern) maps directly onto specific design actions, with each regime identifying which of the four perceptual conditions—shared interpretability, directional monotonicity, local separability, or operational recognisability range—is or is not satisfied.

- R1 (stable gradients) indicate that all four conditions hold for the tested series. The series can be used with reasonable confidence as a graded reference for design exploration up to the first NA level. Because the present study tested one curated series per pair, re-validation is advisable when the series is regenerated with different seeds, prompts or models.
- R2 (stable with weak segments) indicate that the global gradient remains interpretable but at least one local segment fails to satisfy adjacent-level separability. These cases require targeted recalibration rather than full reformulation. Practical responses include re-spacing the intensity levels around the ambiguous segment, reinforcing the perceptual cues that drive the affected transitions through prompt or parameter adjustment, or replacing the problematic levels with regenerated alternatives. Only the localised region needs re-generation and re-testing; the rest of the gradient remains valid.
- R3 (unstable gradients) indicate that the tested configuration does not yield a reliable perceptual ordering. Because regime behaviour is a property of the combined concept–category–pipeline–observer system rather than of any single component, R3 cases admit several distinct remediation paths: (i) reformulating the concept through alternative semantic anchors or related conceptual labels; (ii) modifying the generative pipeline through prompt revision, sampler or scheduler changes, alternative model checkpoints, or LoRA/adapter conditioning to encode the concept more coherently; (iii) changing the product category or base image when the structural constraints of the chosen product interact poorly with the intended manipulation; or (iv) replacing the conceptual direction with a more legible alternative when none of the above produces a stable gradient. R3 thus signals a need

for system-level revision before downstream use, not necessarily a definitive failure of the concept itself.

Across all regimes, NA responses delimit the operational boundary of the axis: once recognisability collapses, further intensification becomes impractical regardless of directional clarity. This framework evaluates the interpretability of the generated series, not aesthetic preference or end-user reception—both of which are subsequent validation stages addressed in Section 5.6.

Finally, once a configuration has been validated, later iterations could rely on smaller panels focused on weak segments or boundary regions; whether such micro-panels preserve diagnostic accuracy remains to be tested.

What this is (and what it is not)

This work does not claim a universal "model of AI understanding". Rather, it proposes a diagnostic frame—R1–R2–R3 × NA—for characterising how AI-driven manipulations of product form manifest perceptually as interpreted by human observers. The novelty lies in:

Reframing the evaluation of AI-generated concept series in terms of perceptual legibility, local separability and recognisable range,

Showing regularities (the three regimes) that are testable beyond our nine cases, and

Making recognisability breakdown an explicit design constraint rather than an artefact to discard.

Reframed in these terms, the present work proposes four perceptual conditions that a generated concept series should satisfy before it is relied upon in design exploration:

- **Shared interpretability** — observers must converge on a consistent ordering of the generated variants, indicating that the target concept evokes a sufficiently shared perceptual interpretation across the relevant audience.
- **Directional monotonicity** — the aggregated ordering must align with the intended direction of intensification, confirming that the generative manipulation produces movement along the concept axis rather than unstructured variation.
- **Local separability** — adjacent intended levels must be perceptually differentiable, so that the gradient supports graded control rather than collapsing into a small number of indistinguishable bands.
- **Operational recognisability range** — concept intensification must remain within the bounds of product-category identity, so that the series remains usable for comparison within the category rather than dissolving into category-loss territory.

When all four conditions hold for a given product–concept–pipeline–observer configuration, the generated series is perceived as a coherent gradient of the intended concept; when one or more fail, the corresponding regime (R2, R3) and the NA boundary identify which condition has not been satisfied. Satisfying these conditions for one curated series is a necessary, but not sufficient, step towards treating a concept as a controllable design variable: control would additionally require showing that the same prompt and parameter settings reproduce the perceived gradient across regenerations, which is left for future work.

The unidimensional treatment of each target concept adopted here is a deliberate first-order operationalisation, appropriate for testing whether a single conceptual direction is perceived consistently. Multi-cue and multi-axis conceptual constructs—where several semantic attributes co-vary or interact—represent a natural extension of this first-order case and are identified as future work.

Limitations

A further limitation concerns the evaluator population. The proposed validation layer is a professional diagnostic tool for early-stage generative design: its purpose is not to assess consumer preference, market acceptance, or the overall success of a final product, but to help design professionals determine whether an AI-generated concept-intensity gradient is interpretable, monotonic, locally separable, and recognisable enough to support design exploration. In this sense, the relevant users of the tool are design professionals rather than the future consumers of the product. This scope condition does not remove the need for later user-based validation with target audiences, nor for studies assessing whether the tool improves design outcomes in practice; it situates the present contribution as a diagnostic step that precedes those stages rather than replacing them. The limitations discussed in the remainder of this section should be read within this intended-use scope.

A key limitation of the present study is its ecological validity. The perceptual judgements were made based on the use of static, uniformly presented AI-generated images. While this controlled presentation isolates formal cues, it does not capture how products are interpreted in real contexts involving materiality, scale, motion, interaction, ambient lighting or social setting. As such, the perceptual gradients observed here should be understood as formal legibility under controlled visual conditions rather than as full accounts of in-situ interpretation.

A further limitation concerns stimulus curation. The base photographs and the representative image for each level were selected by the authors, and one of the curation criteria—consistency with the intended level of concept intensity—entails a prior judgement of the very attribute that observers were asked to rank. This introduces a risk of experimenter bias. Two features of the data mitigate, but do not eliminate, this risk: curation did not guarantee a coherent outcome, since two of the nine curated series produced no reliable ordering; and

observers flagged loss of category recognisability at upper levels that the curation had retained as valid instances of the product. Moreover, each level was represented by a single curated image generated with a fixed seed, so the present results characterise the perceptual structure of these specific series rather than the expected behaviour of the generative configuration across regenerations. Similarly, regime assignment was a qualitative judgement by the authors; Table 1 documents the indicator values on which it rests. Blind curation by independent designers and the evaluation of several seeds per level would allow these sources of bias to be quantified in future studies.

The participant sample consisted exclusively of experienced design engineers, whose training enables fine-grained and calibrated readings of product-form cues. This expertise is appropriate for assessing the interpretability of concept gradients in a design workflow, but it may not represent how non-experts or general consumers would perceive the same variations. Consequently, the stability or instability observed for some gradients may differ in broader or more heterogeneous user populations.

Several of the concepts used in this study—such as “futuristic,” “silence,” or “origami”—have culture-dependent interpretations. Although we selected terms that are widespread in design discourse and likely well represented in the training data of contemporary generative models, their perceptual meaning may vary across cultural contexts, user populations or design traditions. As such, the observed perceptual gradients should be interpreted within the cultural and disciplinary frame of the participant sample.

Although the nine Product–Concept pairs span heterogeneous categories and concept types, they do not exhaust the combinatorial space of possible semantic manipulations—and no finite sample could. The regimes observed in this study should therefore be interpreted as empirical regularities rather than universal patterns, with the understanding that additional

categories, concepts and generative strategies may reveal further variants of gradient behaviour

Conclusions

This study introduced a ranking-based human validation layer designed to assess whether AI-generated conceptual gradients are perceptually coherent—that is, whether observers consistently interpret the intended direction of concept intensification across a series of product-form stimuli. The approach was evaluated through an online study with 26 qualified design engineers across nine Product–Concept pairs spanning diverse product categories and concept types.

Inter-observer agreement varied markedly across pairs (RQ1). Five pairs—Bottle–Origami, Vase–Ruby, Lamp–Polygonal, Table–Contrast and Sneakers–Futuristic—showed strong agreement and recovered the intended A–F order (RQ2); Chair–Aerodynamic and Chair–Aerodynamic 2 showed moderate agreement with local ambiguity between adjacent levels; and Bottle–Bubble and Sneaker–Silence yielded no reliable ordering, so no directional conclusion could be drawn for them. NA responses concentrated at higher intended levels in most pairs and were highly consistent across observers (RQ3), indicating that loss of category recognisability is a shared perceptual boundary rather than an idiosyncratic judgement.

These findings converge into a diagnostic framework of three perceptual regimes—R1 (stable), R2 (stable with weak segments) and R3 (unstable)—combined with the NA-based identification of the recognisable range. For design teams, R1 series can serve as graded references for exploration within that range, R2 series call for local refinement at ambiguous transitions, and R3 series call for revision of the concept, prompt, pipeline or product before use.

The methodological contribution lies in showing that a lightweight protocol based on ranking with explicit recognisability marking and ordinal–probabilistic analysis can characterise the

perceptual structure of concept-driven generative manipulations, a question not directly addressed by existing product-semantics, Kansei or generative-AI evaluation methods. The regime taxonomy offers a vocabulary for describing gradient behaviour across concept–category combinations, and treating recognisability loss as evidence rather than noise makes the usable span of a generated series explicit.

By shifting the focus from output volume to interpretable structure, this work positions human perceptual validation as a necessary step in AI-driven generative design workflows. Establishing whether a perceived gradient can be reproduced reliably across regenerations—and hence whether a concept can be treated as a controllable design variable—is the natural next step.

These conclusions should be interpreted in light of the study's scope. The participant sample was culturally and linguistically homogeneous, the stimuli were static AI-generated images presented under controlled visual conditions, and the nine Product–Concept pairs, while heterogeneous, represent a finite subset of the possible concept–category space. Future work should pursue several complementary directions. First, the generalisability of the regime framework should be tested across broader and more diverse observer populations—including non-expert end-users and cross-cultural samples—as well as under richer stimulus conditions involving materiality, scale, or interactive presentation, extending validation beyond the methodological level addressed here. Second, the protocol should be extended to alternative generative architectures and to concepts drawn from other design traditions to further test the robustness of the R1–R2–R3 taxonomy. Third, causal validation studies that systematically vary individual components of the concept–category–pipeline–observer system—through multiple seeds per level, blind stimulus curation, multi-model comparisons, prompt-controlled experiments or non-AI stimulus baselines—would help disentangle the relative contribution of each factor to regime behaviour. Fourth, the framework should be

extended to multi-concept and multi-axis design scenarios, in which several conceptual attributes co-vary or interact during intensification; the single-axis case addressed here represents a necessary first-order operationalisation upon which such extensions can build. Finally, longitudinal studies examining how calibrated gradients perform across iterative design cycles, and whether micro-panel strategies can effectively reduce validation cost without compromising diagnostic accuracy, constitute promising directions for integrating this approach into routine generative design practice.

References

1. Krippendorff, K. The Semantic Turn: A New Foundation for Design (CRC Press, 2006).
2. Kazmierczak, E. T. Design as meaning making: from making things to the design of thinking. Des. Issues 19, 45–59 (2003).
3. Osgood, C. E., Suci, G. J. & Tannenbaum, P. H. The Measurement of Meaning (University of Illinois Press, 1957).
4. Schütte, S. Engineering Emotional Values in Product Design—Kansei Engineering in Development. PhD thesis, Linköping Univ. (2005).
5. Siddharth, L., Blessing, L. & Luo, J. Natural language processing in-and-for design research. Des. Sci. 8, e16 (2022).
6. Feng, X. et al. Crafting user-centric prompts for UI generations based on Kansei engineering and knowledge graph. Adv. Eng. Inform. 65, 103217 (2025).
7. Guo, F., Qu, Q. X., Nagamachi, M. & Duffy, V. G. A proposal of the event-related potential method to effectively identify Kansei words for assessing product design features in Kansei engineering research. Int. J. Ind. Ergon. 75, 102891 (2020).
8. Perez-Ortiz, M. & Mantiuk, R. K. A practical guide and software for analysing pairwise comparison experiments. Preprint at https://arxiv.org/abs/1712.03686 (2017).

9. Hoeijmakers, E. J. I. et al. How subjective CT image quality assessment becomes surprisingly reliable: pairwise comparisons instead of Likert scale. Eur. Radiol. 34, 4494–4503 (2024).
10. Luce, R. D. Individual Choice Behavior: A Theoretical Analysis (Wiley, 1959).
11. Plackett, R. L. The analysis of permutations. J. R. Stat. Soc. C 24, 193–202 (1975).
12. Turner, H. L., van Etten, J., Firth, D. & Kosmidis, I. Modelling rankings in R: the PlackettLuce package. Comput. Stat. 35, 1027–1054 (2020).
13. Chen, L. et al. AutoSpark: supporting automobile appearance design ideation with Kansei engineering and generative AI. ACM Digital Library (2024).
14. Marrone, A. Optimizing Product Development and Innovation Processes with Artificial Intelligence. Thesis, Politecnico di Torino (2023).
15. Liang, J. The application of artificial intelligence-assisted technology in cultural and creative product design. Sci. Rep. 14, 31069 (2024).
16. Luo, J. Development of an Artificial-Intelligence-Driven Product Design Evaluation Model Using Multi-Modal Data. Thesis, The Hong Kong Polytechnic Univ. (2025).
17. Amershi, S. et al. Guidelines for human–AI interaction. In Proc. 2019 CHI Conference on Human Factors in Computing Systems (ACM, 2019).
18. Fattah Saleh, H. Enhancing the Industrial Design Process with Generative AI. Thesis, Theseus repository (2025).
19. Wang, Y., Zhang, J., Shen, C., Yu, H. & Luo, S. Generative AI aids product design iteration framework based on personalized style and aesthetic preferences. SSRN preprint (2024).
20. Comte, J. AI vs. Human in the Loop: Prototype Designs and Expert Evaluation of Semiautomated Consistency Checking. Thesis (2022). https://www.diva-portal.org/smash/get/diva2:1696863/FULLTEXT01.pdf
21. Ishihara, S., Nagamachi, M., Schütte, S. & Eklund, J. Affective meaning: the Kansei engineering approach. In Advances in Applied Ergonomics 523–530 (2008).
22. Georgiev, G. V. & Georgiev, D. D. Semantic analysis of engineering design conversations. In Proc. DESIGN 2020 International Design Conference (Cambridge Univ. Press, 2020).
23. Klein, S. A. Measuring, estimating, and understanding the psychometric function: a commentary. Percept. Psychophys. 63, 1421–1455 (2001).
24. García-Pérez, M. A. Adaptive psychophysical methods for nonmonotonic psychometric functions. Atten. Percept. Psychophys. 76, 621–641 (2014).

Funding

The authors received no specific funding for this work.

Author contributions

Manuel Ibáñez Arnal conceived the study, designed and conducted the experiment, collected the data, and wrote and formatted the manuscript. Luis Doménech Ballester and Víctor García Peñas developed the mathematical and statistical aspects of the proposed method and provided overall supervision.

Competing interests

The authors declare no competing interests.

Figures

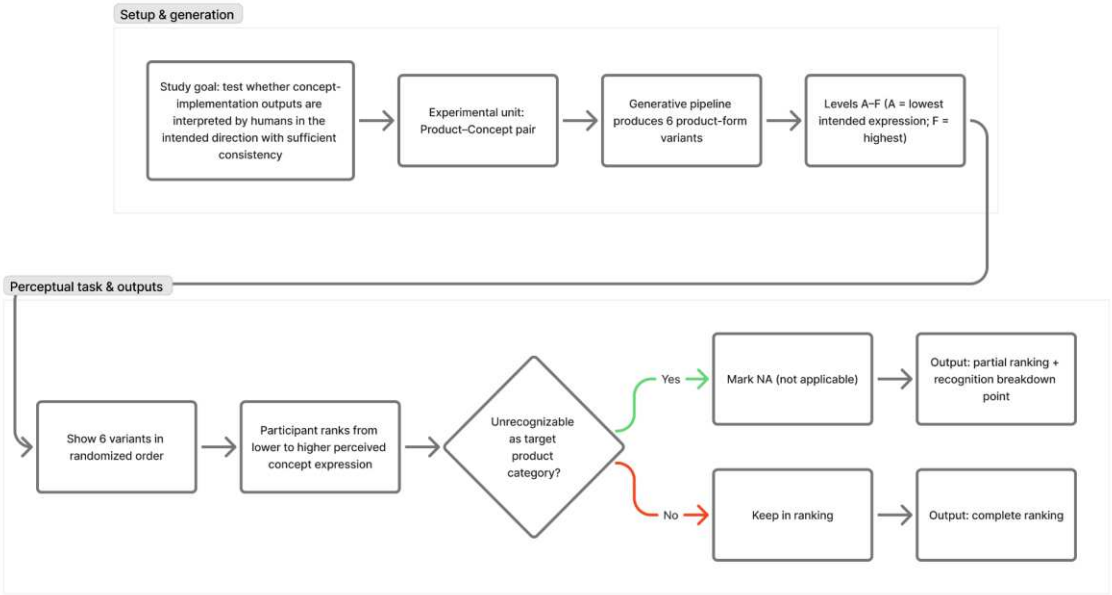


Figure 1: Study protocol and ranking output structure. Top: stimulus setup and generation. Bottom: perceptual ranking task. Product images in this figure are AI-generated stimuli produced with the workflow described in Methods.

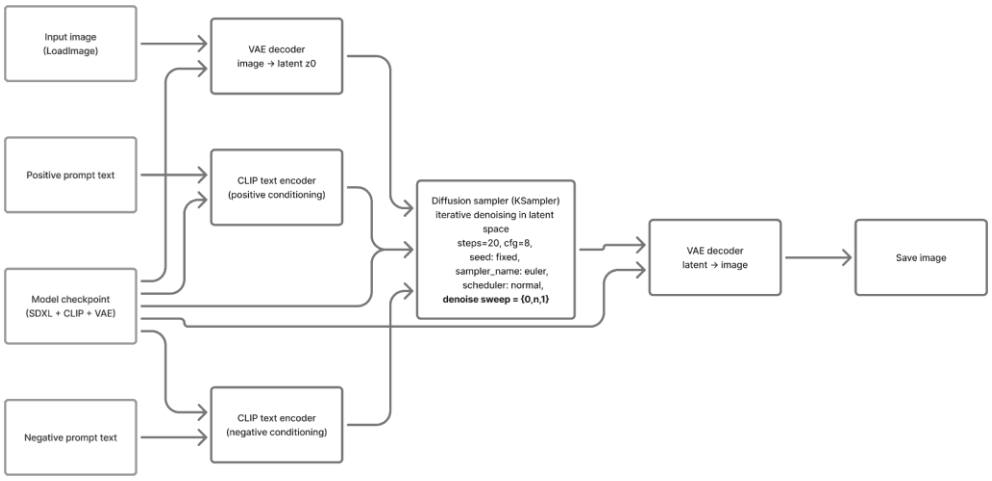


Figure 2: Compact schematic of the latent-diffusion image-to-image workflow

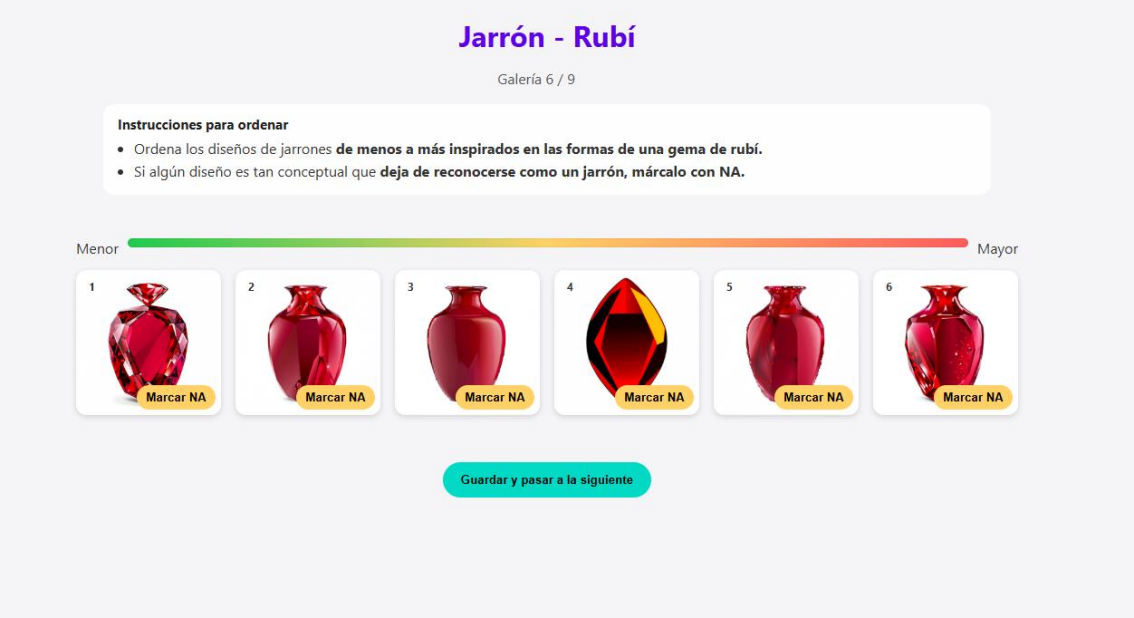


Figure 3: Screenshot of the ranking-based perceptual task interface for the Product–Concept pair “Vase–Ruby”. Six AI-generated product-form stimuli are presented

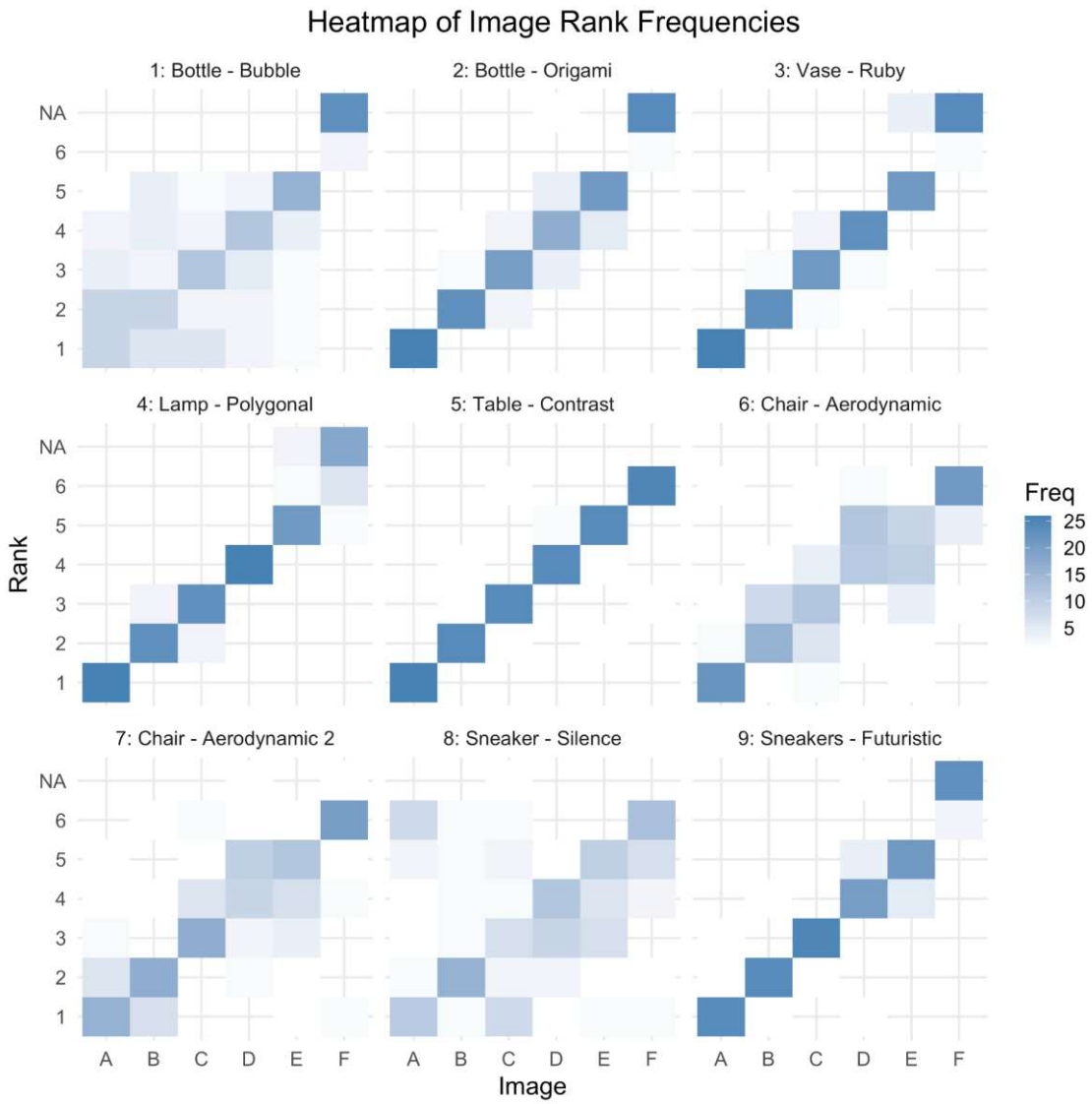


Figure 4: Level–position heatmaps for all nine Product–Concept pairs. Each panel

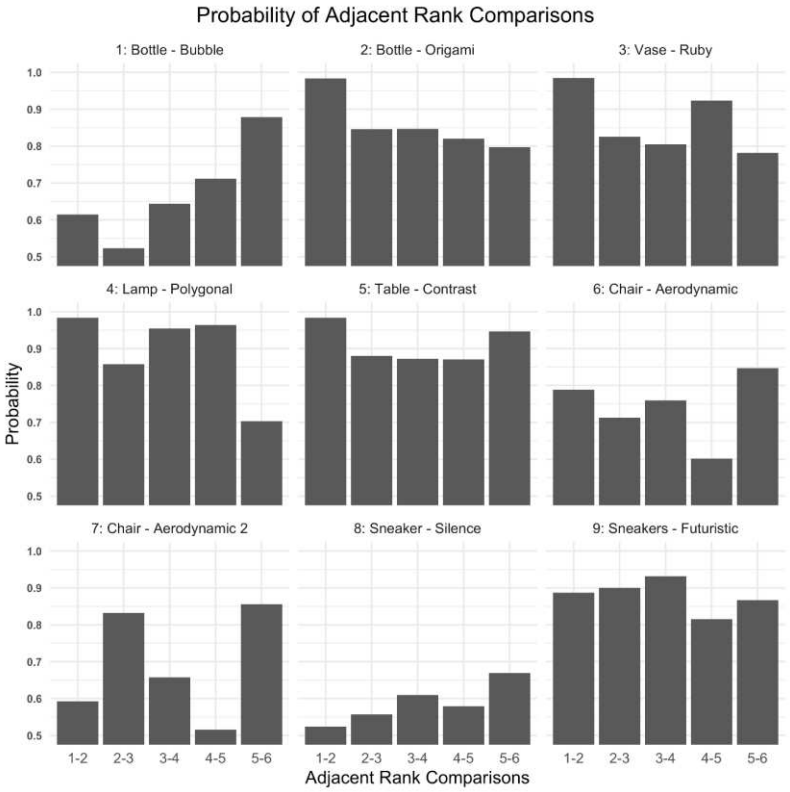


Figure 5: Estimated probability (Plackett–Luce) that the higher-numbered level is preferred over the lower-numbered one.

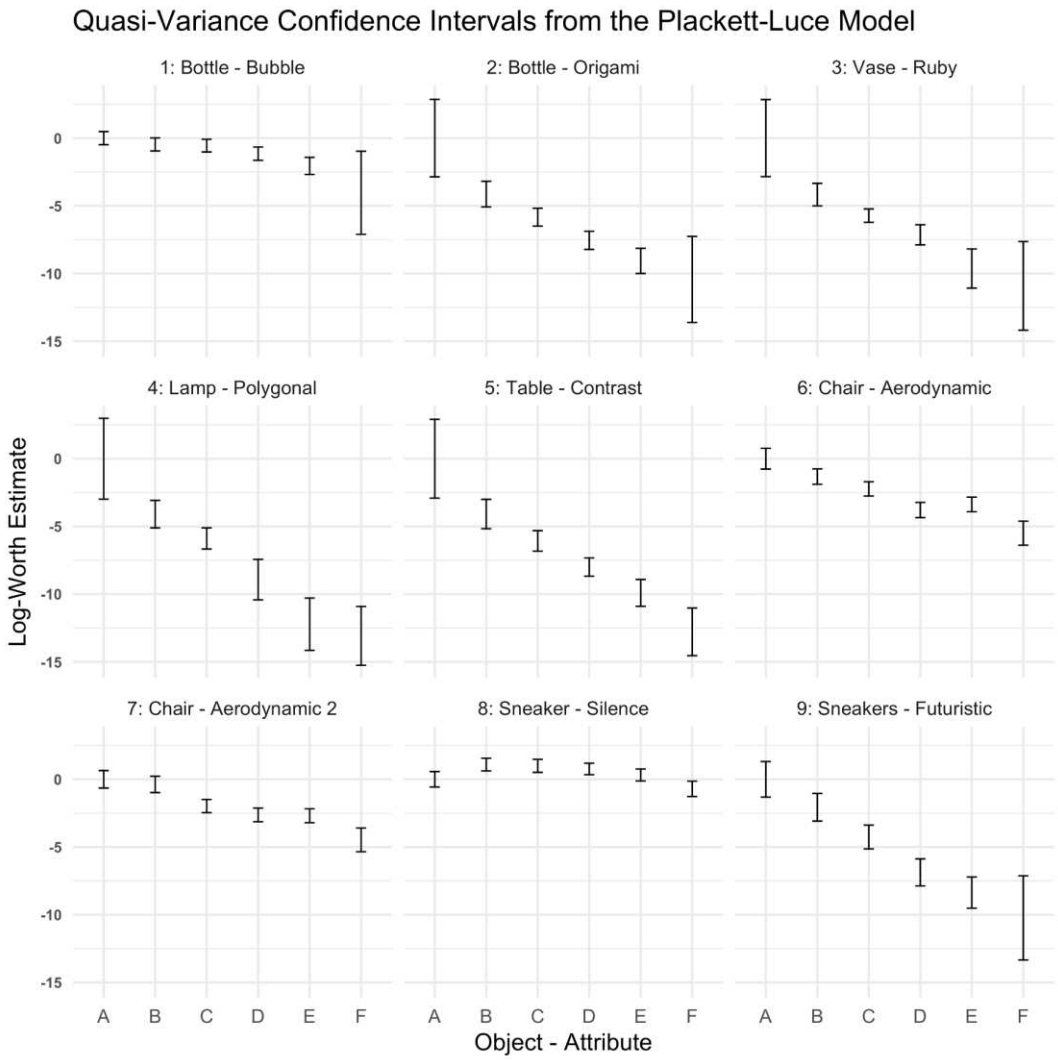


Figure 6: Plackett–Luce log-worth estimates with quasi-variance confidence

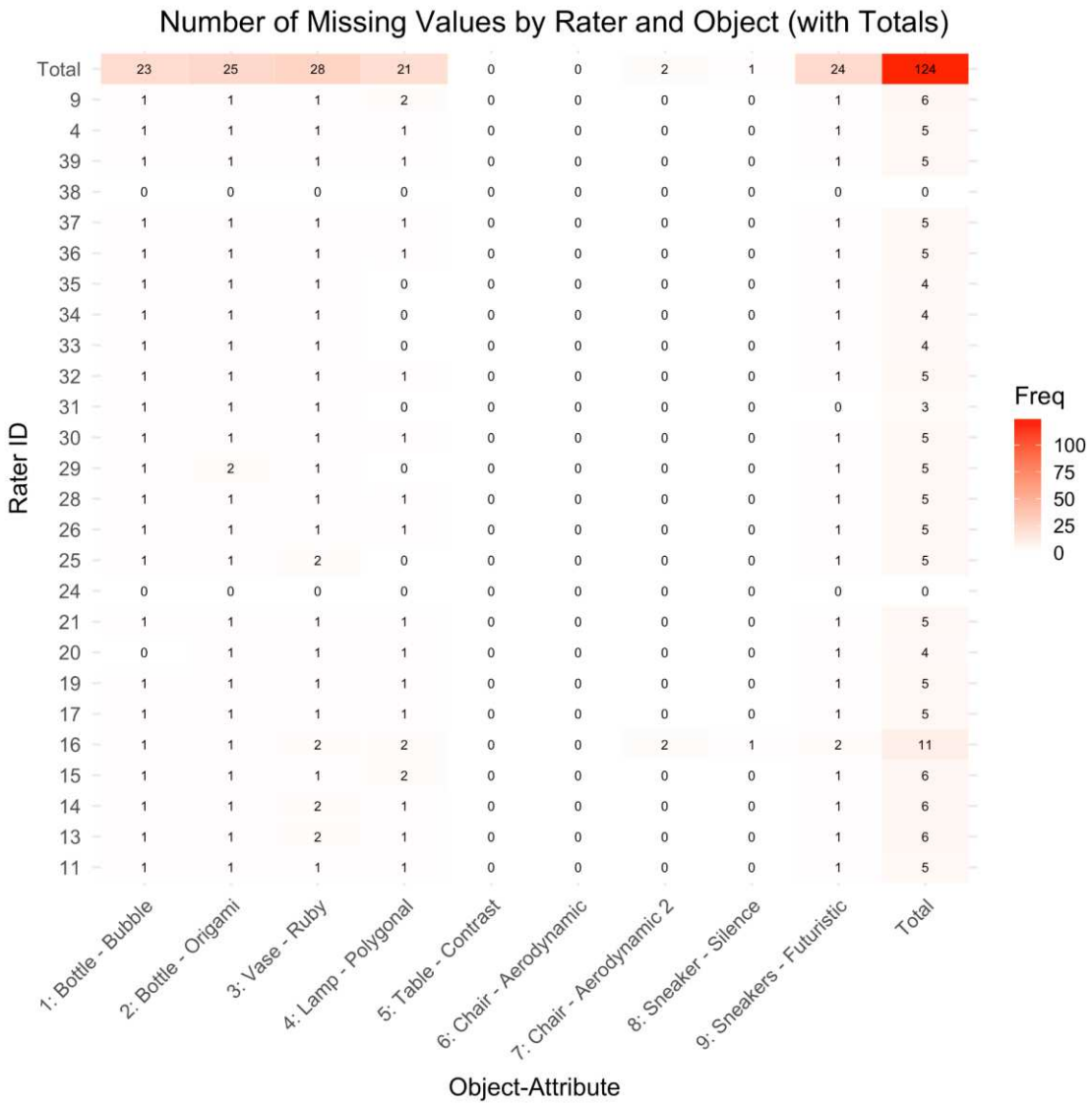


Figure 7: Number of NA responses per observer and Product–Concept pair, with row

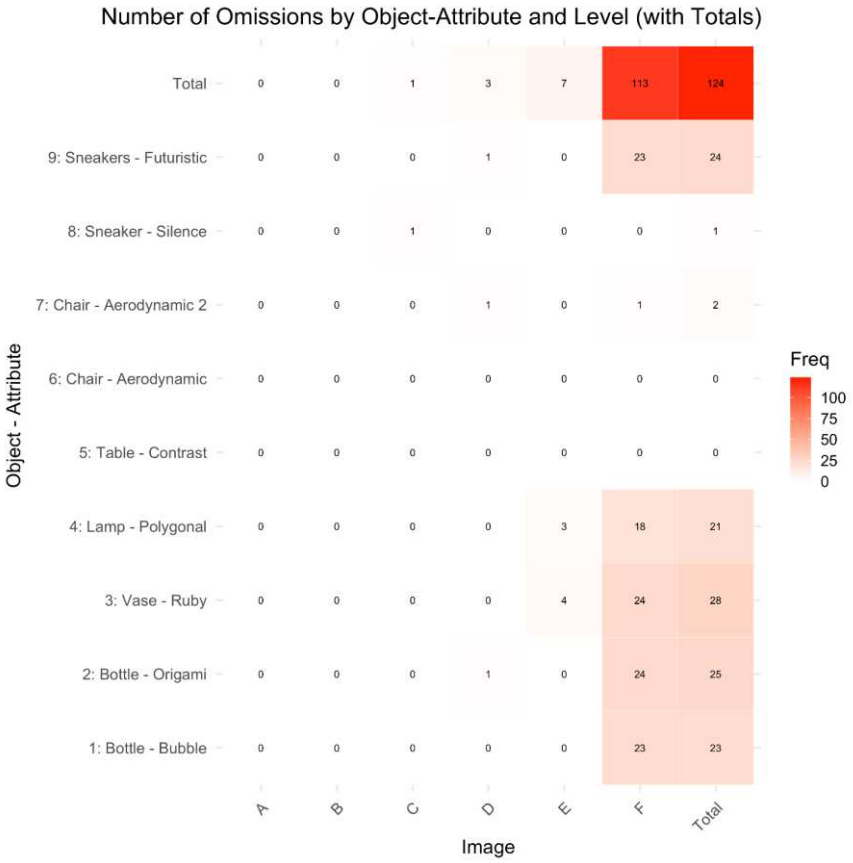


Figure 8: Number of NA responses per intended level (A–F) and

Tables

Regime	Inter-observer agreement (Kendall's W; Krippendorff's α)	Directional alignment (Spearman's ρ)	Local separability (PL adjacent-comparison probabilities)	Heatmap structure
R1 (stable)	$W \approx 0.89–0.98$; $\alpha \approx 0.90–0.98$	$\rho = 1.00$ ($p \leq 0.003$)	Mostly ≥ 0.7; occasional weaker transitions absorbed by high global agreement	Clear diagonal pattern
R2 (stable with weak segments)	$W \approx 0.59–0.72$; $\alpha \approx 0.58–0.71$	$\rho \approx 0.94–1.00$ ($p \leq 0.017$)	One or more transitions noticeably below 0.7	Diagonal pattern with diffuse regions
R3 (unstable)	$W < 0.30$; $\alpha < 0.30$	Not interpretable in the absence of inter-observer agreement	Multiple transitions near 0.5; PL intervals widely overlapping	Dispersed; no clear diagonal

Table 1: Empirical indicator profile of the R1–R2–R3 regimes as observed across the nine Product–Concept pairs in the present study.

Pair	Po	Fleiss' κ	Kripp. α	Gwet AC1	χ^2 p-value	CA p-value	Z_CA	JT p-value	Z_JT
Bottle – Bubble	0.9808	0.8593	0.8593	0.9527	0.00019996	7.44e−09	5.7805740	0.0002	1
Bottle – Origami	0.9808	0.8609	0.8609	0.9488	0.00019996	7.97e−10	6.1454293	0.0002	1
Vase – Ruby	0.9615	0.7632	0.7633	0.9011	0.00019996	5.13e−07	5.0215055	0.0002	1
Lamp – Polygonal	0.9295	0.5312	0.5313	0.8576	0.00019996	1.55e−04	3.7822112	0.0002	1
Table – Contrast	1.0000	NaN	1.0000	1.0000	NaN	1.0	0.0000000	1.0	1
Chair – Aerodynamic	1.0000	NaN	1.0000	1.0000	NaN	1.0	0.0000000	1.0	1
Chair – Aerodynamic 2	0.9872	–0.0130	–0.0127	0.9737	1.0	0.5771	0.5576410	0.1598	1
Sneaker – Silence	0.9936	–0.0065	–0.0062	0.9870	1.0	0.7777	-0.2822883	0.6684	1
Sneaker – Futuristic	0.9744	0.8148	0.8149	0.9348	0.00019996	3.57e−09	5.9030429	0.0002	1

Table 2: Cross-pair summary of NA-based operational-range constraints (nine Product–Concept pairs)

Pair	Kendall's W	Kripp. α	Fleiss κ	Spearman ρ (p)	Kendall τ (p)	Monotonicity	Diagnosis
Bottle–Bubble	0.247	0.268	0.095	1.000 (0.003)	1.000 (0.003)	aggregate only (unreliable)	saturation with low agreement (category drift)
Bottle–Origami	0.914	0.910	0.700	1.000 (0.003)	1.000 (0.003)	yes	saturation / recognisability breakdown at E–F
Vase–Ruby	0.892	0.902	0.742	1.000 (0.003)	1.000 (0.003)	yes	saturation / recognisability breakdown at E–F
Lamp–Polygonal	0.976	0.977	0.858	1.000 (0.003)	1.000 (0.003)	yes	saturation / recognisability breakdown at E–F
Table–Contrast	0.924	0.921	0.865	1.000 (0.003)	1.000 (0.003)	yes	stable monotonic gradient
Chair–Aerodynamic	0.723	0.712	0.350	0.943 (0.017)	0.867 (0.017)	partial	mostly monotonic; minor deviations/ambiguity
Chair–Aerodynamic 2	0.591	0.580	0.271	1.000 (0.003)	1.000 (0.003)	yes	moderate agreement; local ambiguity
Sneaker–Silence	0.203	0.172	0.169	0.429 (0.419)	0.467 (0.272)	no	weak/non-monotonic concept axis
Sneakers–Futuristic	0.898	0.897	0.773	1.000 (0.003)	1.000 (0.003)	yes	saturation / recognisability breakdown at E–F

Table 3: Cross-pair summary of agreement, monotonic alignment (nine Product–Concept pairs)