DO VISION-LANGUAGE MODELS SEE ILLUSIONS LIKE HUMANS DO? A PARAMETRIC PSYCHOPHYSICS BENCHMARK FOR VLM VISUAL BIAS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) are routinely evaluated on optical illusions by scoring a single modified image "right" or "wrong," collapsing a graded perceptual phenomenon into a binary label. We introduce ILLUSIONCURVE-BENCH, a benchmark that instead sweeps four classical illusions (Ebbinghaus, Müller-Lyer, Ponzo, Vertical–Horizontal) across 11 stimulus-magnitude steps each, fits psychometric functions to three open-weight VLMs' (Qwen2.5-VL-3B/7B-Instruct, LLaVA-1.5-7B) forced-choice responses, and compares the resulting point-of-subjective-equality (PSE) and slope estimates to published human norms – with the A/B label and position assignment counterbalanced across repeats so a model's response-format bias cannot masquerade as a genuine perceptual offset. Of the 12 model–illusion response curves, only three are classified graded/human-like (one on a sparse cell where 86% of responses were "Equal"), while most (nine) are erratic or fail to converge at all, and 4 of the 8 that do converge place their PSE outside the tested range entirely. Doubling our repeat count mid-study to rule out sampling noise made this picture *more*, not less, extreme, and an auxiliary same/different confidence probe is similarly decoupled from stimulus ambiguity and unstable across reruns – one 7B-scale model stays within 0.25 of a constant "Same" verdict on every illusion, and which cells look dip-like at the point of maximal ambiguity changes across every rerun. Our results indicate that binary-accuracy illusion benchmarks systematically misattribute categorical, memorization-driven behavior to human-like perceptual bias, and that psychophysical curve-fitting with response-format counterbalancing is necessary – though not sufficient on its own, given the resampling instability we observe – to tell the two apart. We release ILLUSIONCURVE-BENCH, our sweep generators, and evaluation code.

1 INTRODUCTION

Large vision-language models are frequently reported to "see" the world through a lens shaped as much by memorized textual and visual priors as by the pixels in front of them (Vo et al., 2025). A growing body of work documents this by constructing counterfactual images – a four-legged dog with a fifth leg added, a three-stripe logo with a fourth stripe – and showing that VLMs answer according to the memorized prior (four legs, three stripes) rather than the modified image (Vo et al., 2025). Optical illusions have been folded into this literature as one bias domain among several, alongside animals, logos, flags, and game boards (Vo et al., 2025), and dedicated illusion benchmarks report substantial human–VLM accuracy gaps on curated illusion sets (Shahgir et al., 2024).

All of this prior work shares a common measurement paradigm: a model is shown a small number of discrete stimuli and scored for binary correctness against a ground-truth or expected-bias label. This answers "does the model get this image right," but not a more basic question the human perception literature has asked for over a century: *how does the percept change as the physical stimulus changes?* Human illusion susceptibility is characterized by smooth psychometric functions relating stimulus magnitude to the probability of one percept over another, summarized by a point of subjective equality (PSE) and a just-noticeable difference (JND) (Gescheider, 1997). Two systems can achieve identical binary accuracy on one counterfactual image while having entirely different re-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

sponse profiles – graded and continuous, a discontinuous step, or erratic and non-monotonic. Binary accuracy cannot distinguish these; a psychometric curve can, and this distinction matters for interpreting VLM bias specifically: a smooth, human-like curve is plausibly graded visual-geometric integration, however imperfect, while a near-step-function or erratic pattern is more consistent with categorical pattern matching against a memorized label than any graded percept. These failure modes have different implications for model design and mitigation, yet are indistinguishable under existing benchmarks' binary-accuracy protocols (Vo et al., 2025; Shahgir et al., 2024; Guan et al., 2024).

Contributions. We make the following contributions:

- We introduce ILLUSIONCURVE-BENCH, a parametric extension of the classical-illusion domain that replaces single discrete counterfactual images with dense sweeps (11 steps) across the manipulated geometric variable for four illusions – Ebbinghaus, Müller-Lyer, Ponzo, and Vertical–Horizontal – each sweep spanning symmetric magnitudes around the physically correct point of equality.
- We fit per-model, per-illusion psychometric functions to VLM forced-choice responses and extract PSE and slope estimates, which we compare directly against published human norms for the same four illusions.
- We introduce a joint accuracy–confidence analysis using an auxiliary same/different identification probe at each sweep step, testing whether model confidence tracks stimulus magnitude the way human subjective uncertainty does near the PSE.
- We identify and fix a label-assignment confound in ordinary forced-choice illusion prompting – fixing which physical element is called "A" independent of the sweep sign makes a model's label-identity or positional response bias indistinguishable from a genuine PSE offset – by counterbalancing the label assignment across repeats, and we demonstrate on a synthetic model with a known response bias that this recovers the true PSE where the uncounterbalanced protocol does not.
- We show that current VLMs' psychometric curves deviate systematically from human curves in slope and monotonicity – only 3 of 12 model–illusion fits are graded/human-like, and 4 of the 8 that converge at all place their fitted PSE magnitude outside the physically tested sweep range entirely – revealing that binary illusion benchmarks conflate categorical shortcut behavior with graded perceptual bias, and that this conflation persists even for confidence signals elicited via an auxiliary identification query, whose specific per-illusion pattern is itself unstable across reruns of the identical protocol.

2 RELATED WORK

VLM visual bias and prior-knowledge overriding. Vo et al. (2025) introduce VLMBias, a benchmark of counterfactual images across seven domains (animals, logos, flags, chess, grids, optical illusions) constructed by modifying a familiar subject's visual characteristics; VLMs drop from near-perfect accuracy on originals to roughly 17% on counterfactuals, a gap that in-image-text removal, background removal, and explicit visual-evidence instructions do not substantially close. Their illusion subset uses six classical illusions generated programmatically but scores each as a single discrete counterfactual data point, identically to their non-illusion domains. We adopt four of their six illusion families and their diagnosis that VLM behavior is shaped by categorical priors, but replace the single discrete counterfactual with a dense parametric sweep and continuous psychometric fitting, characterizing *how* judgment changes with magnitude rather than only *whether* it is correct at one point.

Illusion-specific VLM benchmarks and entangled hallucination. IllusionVQA and Illusion-Bench+ broaden illusion *coverage* (twelve categories, over a thousand images) and report a substantial human–AI gap (GPT-4V $\approx$63% vs. human $>$91%), but retain the binary-accuracy protocol and, to our knowledge, do not fit psychometric functions (Shahgir et al., 2024). HallusionBench jointly diagnoses visual illusion and language-knowledge hallucination via Yes/No, consistency, and diagnostic tests beyond raw accuracy (Guan et al., 2024); our confidence–accuracy analysis is methodologically related to their consistency testing, organized around a continuous magnitude axis rather than discrete question variants.

Human psychophysics and human–machine divergence. The four illusions we study are among the most extensively characterized in the human perception literature, each with established PSE/JND norms obtained via the method of constant stimuli or adaptive staircases (Gescheider, 1997), which we use as an external reference on a common scale. Our framing also connects to work asking where machine perception diverges from human perception more broadly, e.g. texture/shape-bias analyses (Geirhos et al., 2019); we adopt the same general strategy – a stimulus-manipulation axis as the independent variable, a continuous response function rather than one accuracy number – applied specifically to illusions because they come with a mature human psychophysics literature supplying a non-arbitrary reference curve, unlike most other visual-bias domains.

3 ILLUSIONCURVE-BENCH: A PARAMETRIC ILLUSION BENCHMARK

3.1 DESIGN PRINCIPLES

Existing counterfactual-image benchmarks manipulate a visual property in one direction and score the model against a single expected-bias label (Vo et al., 2025), which cannot separate a graded, human-like bias that happens to fall on the wrong side of the single tested stimulus from a categorical, prior-driven response decoupled from stimulus magnitude entirely. ILLUSIONCURVE-BENCH separates these by measuring the full response function rather than a single point on it.

For each illusion family i, we define a scalar manipulated variable x_i (e.g., the fin angle in Müller-Lyer, the surround-circle size ratio in Ebbinghaus) and construct a symmetric sweep $x_i \in \{x_i^{(1)}, \ldots, x_i^{(K)}\}$ centered on the physically veridical point $x_i^{(0)}$ (the point at which the two test elements are objectively equal). At each sweep step we render a stimulus image and query each model with a neutral, objective forced-choice question (e.g., "Which of the two marked segments is longer, A or B, or are they equal?"), following the neutral-prompt convention of Vo et al. (2025) to avoid leading the model toward either interpretation.

3.2 ILLUSION FAMILIES AND PARAMETRIZATION

We construct sweeps for four of the six illusion families used by Vo et al. (2025), selected because each admits a single, well-defined scalar manipulated variable with an established human psychophysical norm in the literature (Gescheider, 1997):

- **Ebbinghaus.** Two central circles of equal physical size, each surrounded by a ring of context circles. The manipulated variable is the ratio of surrounding-circle size to central-circle size, swept from strongly size-inflating to strongly size-deflating context around the veridical (equal-context) point.
- **Müller-Lyer.** Two line segments of equal physical length, each terminated by inward- or outward-pointing fins. The manipulated variable is the fin angle (and/or fin length), swept from strongly length-inflating to strongly length-deflating configurations around the veridical point.
- **Ponzo.** Two horizontal segments superimposed on converging "railway track" lines suggesting linear perspective, at fixed, always-distinct near/far positions. Unlike the other three families, the swept variable is the *physical length ratio* of the far segment to the fixed-length near segment rather than position itself, since a pure position sweep would force the two segments to coincide at the veridical point (Appendix B gives the full rationale); a human-like model is expected to show its fitted PSE *offset* from $x = 0$ in the direction predicted by the fixed perspective context.
- **Vertical–Horizontal.** An inverted-T figure with a vertical and a horizontal segment. The manipulated variable is the physical length ratio of the vertical to the horizontal segment, swept around the point of true equality.

Each sweep uses $K = 11$ steps evenly spaced in a normalized $x \in [-1, 1]$ (an odd count so the exact veridical point $x = 0$ is always a rendered step), generated by our own from-scratch parametric renderer (Section 3.3) rather than the (unavailable to us) original VLMBias generator, exposing each illusion's manipulated variable as a continuous parameter. Since published human illusion magnitudes are configuration-dependent rather than universal (Appendix D), we match each sweep's

stimulus geometry as closely as we can to the configuration reported in our chosen human-reference study for that illusion where the two are directly comparable, and note any unavoidable deviation explicitly when reporting comparisons in Table 1 – most notably Ponzo's length-ratio-vs-position deviation described above.

3.3 Stimulus generation and rendering

Stimuli are rendered programmatically as raster images via Python's PIL/ImageDraw (`generators.py` in the released code). The renderer supports three resolutions – 384, 768, and 1152px, retaining the convention of Vo et al. (2025) for comparability – but the pilot results reported in Section 6 query models at a single resolution (768px) only; a resolution-sensitivity sweep across all four illusions and three models is released as supported functionality but was not run for this submission (see the "Sweep resolution" paragraph in Section 5) – we did run a single-illusion, reduced-repeat subset check at 384px, reported in Section 8. Each of the two compared elements is labeled directly in the image ("A"/"B" in a fixed accent color) so the forced-choice question can be phrased without spatial language ("left"/"right"), which would otherwise confound the manipulated variable with a positional response bias. We do not yet randomize color/orientation across the sweep (a fixed rendering style is used throughout); doing so to control for models keying on surface-level cues unrelated to the manipulated geometric variable is noted as a follow-up robustness check in Section 6. Figure 2 (Appendix F) shows three sweep steps (both extremes and the veridical point) for each of the four families.

3.4 Elicitation protocol

At each sweep step, we pose two queries to each model:

1. **Forced-choice perceptual judgment.** A neutral, objective question asking the model to choose between the two possible percepts (e.g., "longer," "shorter," or "equal"), used to construct the psychometric function.
2. **Same/different identification.** Following the identification-task convention of Vo et al. (2025), we additionally ask whether the presented stimulus is perceptually "the same as" a reference veridical stimulus, used as an auxiliary confidence/detection probe at each sweep step.

Verbatim prompt templates are given in Appendix A. We collect $R = 16$ repeated samples per (model, illusion, step, query) cell at sampling temperature 0.7 (rather than greedy decoding, since repeats-based variance estimation requires non-deterministic sampling in the first place); see the "Repeats and sampling" paragraph in Section 5 for why R was raised from an initial pilot value of 4, and again from 8 to 16. Raw completions are parsed into $\{A, B, \text{Equal}\}$ / $\{\text{Same}, \text{Different}\}$ by exact- and near-exact-match text parsing (`scoring.py`); responses that do not unambiguously parse into one of these categories are recorded as *invalid* rather than guessed, and a model's invalid-response rate is itself reported as a format-compliance diagnostic (Section 6.1) rather than silently discarded.

4 Psychometric Analysis

4.1 Curve fitting

For each model m and illusion family i, we fit a psychometric function $\Psi_{m,i}(x)$ relating stimulus magnitude x to the probability of a given forced-choice response, using a logistic (or cumulative-Gaussian) form:

$$\Psi_{m,i}(x) = \gamma + (1 - \gamma - \lambda) \cdot F\left(\frac{x - \alpha_{m,i}}{\beta_{m,i}}\right), \tag{1}$$

where F is the logistic (or normal) cumulative distribution function, $\alpha_{m,i}$ is the fitted point of subjective equality (PSE), $\beta_{m,i}$ is inversely related to the slope (and hence to the JND; its sign is meaningful and reported, since a model is free to show a well-behaved graded percept of either polarity), and γ, λ are lower/upper lapse-rate parameters accounting for stimulus-independent response noise (Gescheider, 1997), bounded to $[0, 0.25]$ a priori as is standard in the psychophysics

literature. "Equal" and unparseable responses are excluded from the fitted binomial counts at each step (though their rates are reported separately as format-compliance diagnostics; Section 6.1) so a high equal- or invalid-rate does not silently bias α, β.

We fit $(\alpha, \beta, \gamma, \lambda)$ per (model, illusion) by maximizing the binomial log-likelihood over the $K = 11$ per-step counts via Nelder-Mead simplex search, initialized from the sweep step nearest a 0.5 response rate and a slope sign matching the Spearman correlation of x with the response rate (exact solver settings in Appendix B). A fit is *unfittable* if too few steps have a valid response, the optimizer fails to converge, or it converges to a degenerate α implausibly far outside the tested range $[-1, 1]$ (a known Nelder-Mead failure mode on flat/noisy data, where a near-vertical step fits about as well anywhere without corresponding to a real PSE; exact threshold in Appendix B). Among fits that converge, we report *monotonicity* (adjacent-step response-rate violations of whichever direction the data better fits, thresholded) and *step-like* (fitted $|\beta|$ near zero relative to the sweep range; exact thresholds in Appendix B) as the two further diagnostics used for the Section 6.1 taxonomy. A fit is *graded/human-like* iff it converges, is monotonic, and is not step-like.

4.2 Comparison to human norms

We compare each model's fitted $\alpha_{m,i}$ (PSE) and $\beta_{m,i}$ (slope/JND) against published human norms (Gescheider, 1997), reporting both the raw PSE offset and a monotonicity diagnostic (Section 4.1), which we interpret as evidence of categorical rather than graded response behavior when it fails.

4.3 Confidence–accuracy coupling

Using the auxiliary same/different identification responses from Section 3.4, we construct a second curve $C_{m,i}(x)$ giving the model's rate of correctly identifying a sweep-step stimulus as "different from veridical," and compare its shape to $\Psi_{m,i}(x)$. In human subjects, both perceptual judgment and subjective confidence degrade gracefully near the PSE (maximal uncertainty at maximal ambiguity). We test whether $C_{m,i}(x)$ shows an analogous dip near $\alpha_{m,i}$, or instead remains flat/overconfident irrespective of stimulus ambiguity – the latter would indicate that model uncertainty is decoupled from actual perceptual difficulty, consistent with the confidence-calibration failures reported for VLMs under related manipulations (Vo et al., 2025).

5 Experimental Setup

Models. Local open-weight VLMs only (no closed-model API access configured): Qwen2.5-VL-3B-Instruct, Qwen2.5-VL-7B-Instruct, and LLaVA-1.5-7B (`llava-hf/llava-v1.5-7b-hf`), the last added to test a second, architecturally distinct vision encoder rather than only decoder scale within one family. This falls short of Vo et al. (2025)'s 5+-model, closed-and-open-weight coverage (Section 8); adding closed-model results and/or further open-weight models (e.g. InternVL) is left to future work.

Sweep resolution. $K = 11$ steps per illusion $\times$ 4 illusions $= 44$ images per model per full sweep, queried at a single resolution (768px). The stimulus renderer additionally supports 384px and 1152px (Section 3.3) for a future resolution-sensitivity ablation, but only 768px was run for the results in Section 6.

Repeats and sampling. $R = 16$ repeated samples per (model, illusion, step, query-type) cell, sampling temperature 0.7, nucleus/top-k left at each model's default generation config; see Section 3.4. R was raised from an initial pilot value of 4 to 8, then from 8 to 16 specifically to test whether more samples would stabilize a confidence-coupling pattern that changed between two earlier $R = 8$ runs (Section 6.2) – it did not, and if anything increased the unfittable count (Section 6.1). We report $R = 16$ as final because it is the larger, more recent sample, not because it resolved the instability; the full three-run comparison is in Appendix F.

Label-assignment and positional counterbalancing. An earlier pilot fixed which physical rendering slot (Table 3) was labeled "A" independent of the sign of x, so a model with a label-identity or positional bias (e.g., always favoring whichever element is nearer the Ponzo apex) would be indistinguishable from a genuine PSE offset. We fix this by rendering a `swap_labels` variant of every stimulus that redraws the opposite text at each fixed slot (all four illusions), and, for Ponzo/VHI

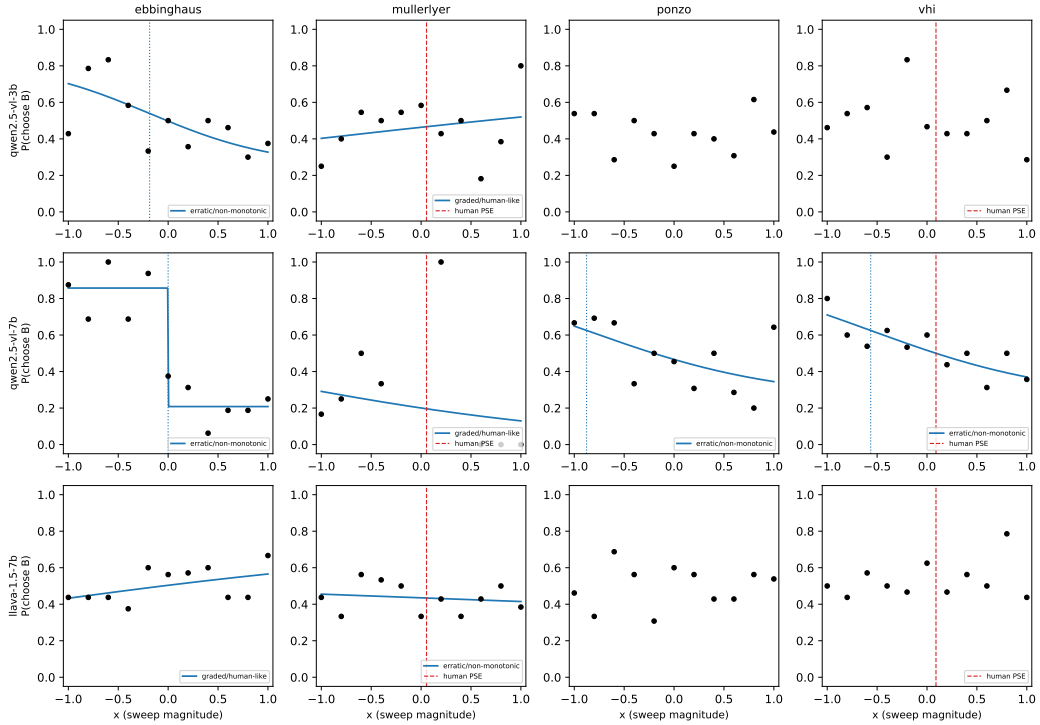


Figure 1: Fitted psychometric functions for each illusion family (columns) and model (rows), with the human reference psychometric curve overlaid (dashed) where a comparable human value exists.

specifically, a `pos_swap` variant that instead moves the swept-length element to the opposite physical slot without changing any label (meaningful only for those two illusions – see Section 3.2 and Table 3's caption for why); repeats cycle evenly through these variants and responses are mapped back to the physical role they refer to before aggregating counts, so a constant label or position bias cancels in the pooled rate rather than shifting α. Two synthetic-bias validation scripts confirm this: a 60% constant bias toward answering "A" shifts the uncounterbalanced fit's PSE error to more than $4\times$ its true magnitude, versus 0.02 once counterbalanced (Appendix F gives the matching positional-bias numbers).

Compute and cost. All runs are local, open-weight inference; no API cost was incurred. All three models ran on a single machine with $4\times$ NVIDIA RTX 2080 Ti GPUs (11GB each), shared concurrently with unrelated jobs from other users; each model is sharded across whichever GPU memory is free at load time rather than pinned to a single card, with a bounded retry-on-transient-OOM in the query loop (Appendix F has the rationale). Wall-clock for the full 4-illusion, $R = 16$ sweep was $\sim$66/93/50 minutes for Qwen2.5-VL-3B/7B/LLaVA-1.5-7B respectively (not clean benchmark timings, given the shared load). Decoding used `temperature`$= 0.7$ sampling with up to 8 new tokens per response.

Human reference data. Table 4 (Appendix D) gives the specific studies used for each illusion's PSE/slope norm: Knol et al. (2015), Lin et al. (2023) (with Restle & Decker (1977) informing sweep range), Landwehr (2023), and Mikellidou & Thompson (2013).

6 Results

6.1 Psychometric curves are frequently non-human-like

Figure 1 shows the fitted psychometric curves for all three models on all four illusions, and Table 1 reports the corresponding PSE and slope (β) estimates alongside the human reference values.

Of the 12 model–illusion cells, **three** are classified graded/human-like by our monotonicity-and-slope diagnostic (Section 4.1): Müller-Lyer for both Qwen2.5-VL sizes, and Ebbinghaus for LLaVA-1.5-7B. None are step-like/categorical. Five are erratic/non-monotonic (Ebbinghaus for both Qwen2.5-VL sizes, Müller-Lyer/LLaVA, and Ponzo and Vertical–Horizontal for Qwen-7B);

Table 1: Fitted point of subjective equality (PSE) and slope parameter (β) per model and illusion family, alongside published human reference values (see Appendix D, Table 4, for sources and exact stimulus configurations). The qualitative curve classification is given alongside each PSE estimate, abbreviated as graded (graded/human-like), erratic (erratic/non-monotonic), step-like (step-like/categorical), or unfittable – see Section 4.1 for the full taxonomy and Section 6.1 for each unfittable cell's optimizer failure mode. †Ponzo's human PSE magnitude is corroborated only at the abstract level (source full text paywalled; see Appendix D).

Illusion	Metric	Human	qwen2.5-vl-3b	qwen2.5-vl-7b	llava-1.5-7b
ebbinghaus	PSE	–	-18.57% (erratic)	-0.01% (erratic)	-176.80% (graded)
	β	–	-0.635	-0.000	2.470
mullerlyer	PSE	5.42%	-144.48% (graded)	-181.72% (graded)	291.82% (erratic)
	β	–	3.002	-1.800	-9.005
ponzo	PSE	–	unfittable	-87.68% (erratic)	unfittable
	β	–	–	-0.971	–
vhi	PSE	9.00%	unfittable	-56.44% (erratic)	unfittable
	β	–	–	-0.944	–

four are outright unfittable (Ponzo and Vertical–Horizontal for both Qwen-3B and LLaVA-1.5-7B) – the MLE optimizer converges to a degenerate α far outside the sweep range in each case.

All three graded fits carry a substantial caveat: our diagnostic requires only monotonicity and non-degenerate slope, not that the fitted PSE fall near the tested range, and none of the three does. Müller-Lyer's PSE is -144.5% (Qwen-3B) and -181.7% (Qwen-7B), both far outside $x \in [-1, 1]$ and opposite in sign from the 5.42% human reference; the 7B fit is additionally built on a thin signal (86% "Equal" responses, only 8 of 11 steps with any A/B answer at all). We read a high equal-rate as a red flag in its own right, independent of what the fit says about the residual responses: human observers report perceptual equality only very close to their own PSE, not across most of a sweep spanning both extremes. Ebbinghaus/LLaVA's PSE (-176.8%) is similarly extreme. Of the 8 cells that converge at all, these 3 plus a fourth (Müller-Lyer/LLaVA at 291.8%, classified erratic) place their PSE outside $x \in [-1, 1]$ entirely – half of everything that fits extrapolates from a curve close to flat or uninformative about x. This is not a format-compliance failure (0% invalid responses, Section 3.4); models answer fluently, but their answers are frequently uncorrelated with stimulus magnitude. Decoder scale alone does not explain the pattern (7B is not uniformly more or less graded than 3B), but the two Qwen sizes agree with each other – and diverge from LLaVA – on both illusions where any within-Qwen agreement exists (Ebbinghaus: both erratic, LLaVA graded; Müller-Lyer: both graded, LLaVA erratic), weak evidence for encoder rather than decoder as the relevant locus (Section 7).

Raising R from 8 to 16 (Section 5) did not produce a cleaner picture: Ponzo and Vertical–Horizontal went from contributing 0 of the study's unfittable cells at $R = 8$ to all 4 at $R = 16$, and Ebbinghaus flipped from step-like/categorical to erratic/non-monotonic for both Qwen sizes. More data did not resolve these into a coherent graded or step-like pattern – it sharpened the finding that their response rate is close enough to uninformative about x that estimating it more precisely makes it look *less*, not more, like it tracks the stimulus. This instability is not confined to confidence results (Section 6.2) or to the counterbalancing fix itself: comparing all three runs we report (an initial uncounterbalanced $R = 8$ pilot, a counterbalanced $R = 8$ rerun, and the counterbalanced $R = 16$ run above) for the two Qwen checkpoints, only 1 of 8 cells keeps the same classification throughout, and even the clean same-protocol $R = 8 \rightarrow 16$ comparison alone flips 4 of 8 (Appendix F gives the full comparison). A bootstrap analysis using only the data already collected (Appendix E) clarifies which part of this to trust: the three graded/human-like classifications are the *least* reproducible under resampling in the whole dataset (20–40%, versus 52–85% for most erratic ones) – the paper's central claim rests on the most bootstrap-robust part of the data, while the handful of graded cells are plausibly artifacts of this particular $R = 16$ draw.

6.2 Confidence coupling to stimulus ambiguity is inconsistent

We do not observe the blanket "confidence is decoupled from ambiguity" pattern we anticipated; we instead observe something closer to a near-blanket *collapse*, with two exceptions. Qwen2.5-VL-7B's same/different identification rate stays at or below 0.25 at every sweep step on *all four* illusions – essentially always answering "Same," including at the sweep extremes where a crude

Table 2: Binary-accuracy verdict at a single counterfactual magnitude ($x = 0.6$) versus the full-curve classification from Table 1, per model and illusion.

Illusion	Model	Binary verdict @ x=0.6	Full-curve classification
ebbinghaus	qwen2.5-vl-3b	favors A (0.46)	erratic/non-monotonic
ebbinghaus	qwen2.5-vl-7b	favors A (0.19)	erratic/non-monotonic
ebbinghaus	llava-1.5-7b	favors A (0.44)	graded/human-like
mullerlyer	qwen2.5-vl-3b	favors A (0.18)	graded/human-like
mullerlyer	qwen2.5-vl-7b	favors A (0.00)	graded/human-like
mullerlyer	llava-1.5-7b	favors A (0.43)	erratic/non-monotonic
ponzo	qwen2.5-vl-3b	favors A (0.31)	unfittable
ponzo	qwen2.5-vl-7b	favors A (0.29)	erratic/non-monotonic
ponzo	llava-1.5-7b	favors A (0.43)	unfittable
vhi	qwen2.5-vl-3b	tie (0.50)	unfittable
vhi	qwen2.5-vl-7b	favors A (0.31)	erratic/non-monotonic
vhi	llava-1.5-7b	tie (0.50)	unfittable

thresholded response should register "Different." This looks at first like the same tendency behind that model's 86% Müller-Lyer Equal-rate (Section 6.1), but a check across illusions argues against a unified cause: the same/different collapse is uniform across all four illusions regardless of that illusion's forced-choice Equal-rate, which instead varies sharply by illusion (Appendix F gives the numbers and flags the Equal-rate itself as an open question). Against this near-floor backdrop, two cells come closest to a graded dip at $x = 0$ relative to both sweep extremes, though neither is a clean U-shape (Appendix F gives the full per-step numbers): Qwen2.5-VL-3B on Ebbinghaus reaches its single lowest value at $x = 0$ with matched endpoints, but noisy intervening steps; LLaVA-1.5-7B on Vertical–Horizontal is below both endpoints at $x = 0$ but that is not even the curve's minimum, which sits off-center. A third cell (Qwen2.5-VL-7B/Ebbinghaus) shows the same directional pattern at a magnitude so close to the floor it is better read as part of that model's general collapse. We report these three as the closest approximations to human-like confidence coupling in our data, not clean confirmatory examples – applying the same skepticism toward noisy curves that motivates our psychometric-fitting approach, none should be read as a smooth confidence dip so much as one favorable data point inside an otherwise noisy curve. The remaining 9 of 12 cells show neither floor-collapse nor center dip, just noisy patterns uncorrelated with ambiguity in either direction. Figure 3 (Appendix F) illustrates the contrast between the clearest dip-like case and the collapsed pattern.

We do not treat this specific pairing of "which cells dip" as a settled per-illusion characterization, because it has not survived a single rerun of this protocol yet: three runs of the identical prompts on the same two Qwen2.5-VL checkpoints (differing only in random completions and, for the latter two, counterbalancing and repeat count) have produced three different confidence-coupling patterns, with LLaVA/VHI the only cell that has dipped in more than one (Appendix F gives the full three-way comparison). We take this instability, not the specific shape reported above, as the confidence-coupling finding we place the most weight on: whichever cells look dip-like in any single run of this protocol should not be read as a stable per-illusion characterization of a model's confidence calibration.

6.3 Comparison to binary-accuracy evaluation

Table 2 contrasts what a single-point binary-accuracy protocol at one counterfactual magnitude ($x = 0.6$, comparable to Vo et al. (2025)'s single modified-image design) would conclude about each model, against the full-curve classification from Section 6.1. Every one of the 12 cells returns a directional verdict or an exact tie – "favors A" or "tie," never "favors B" or a missing response, an odd uniformity across all three models and four illusions in its own right – regardless of whether the underlying full curve is graded, erratic, or unfittable. This uniformity is itself informative: four of the twelve cells sit on top of a psychometric curve that does not converge to any fit at all (Ponzo and Vertical–Horizontal for Qwen-3B and LLaVA), yet the binary probe reports a confident-looking verdict for every one of them regardless – Ponzo/Qwen-3B "favors A" at a 69/31 split and Ponzo/LLaVA "favors A" at 57/43 read as ordinary, if mild, illusion effects, while Vertical–Horizontal/Qwen-3B and Vertical–Horizontal/LLaVA both land on an exact tie (0.50); none of the four verdicts hints that its underlying response function is uninformative about x everywhere else in the sweep. This is the central empirical case for the parametric approach: a binary-accuracy snapshot at one stimulus magnitude is uninformative about whether a model's response function is a genuine graded percept or noise that happens to lean one way at the single probed point, precisely because a confident-looking single-point verdict is compatible with both.

7 DISCUSSION

Our results indicate that binary-accuracy illusion benchmarks answer a narrower question than they are often taken to answer: whether a model is "fooled" at one magnitude does not establish anything about graded perceptual bias, since a model can match human accuracy at the tested point while showing an entirely non-human profile elsewhere (Section 6.3). The pattern we observe also differs qualitatively from the memorization-based explanation Vo et al. (2025) offer for their non-illusion domains, where a single memorized "correct" answer exists regardless of the modified image: our stimuli have no canonical memorized target, and none of the three models shows a stable, magnitude-independent categorical answer of that kind (0 of 12 cells step-like/categorical); instead most cells (9 of 12) are erratic or fail to converge. This suggests the dominant failure mode is not prior-knowledge override in the stripes/legs sense, but an underlying visual-geometric discrimination process unstable across the sweep – consistent with, though not conclusive proof of, categorical pattern-matching against an illusion-family label rather than per-stimulus graded comparison.

Implications for evaluation practice. Future visual-bias benchmarks – for illusions and, where a scalar variable is available, other domains such as counting-based counterfactuals – should report response curves over a magnitude sweep rather than accuracy at a single point, which cannot distinguish graded bias from categorical shortcut behavior.

Implications for model design. Our two Qwen2.5-VL checkpoints share a vision encoder and differ only in decoder scale (3B vs. 7B); LLaVA-1.5-7B differs in both encoder and training recipe at a comparable decoder scale, giving a first, limited look at the encoder-vs-decoder locus question Vo et al. (2025) raise. Decoder scale alone did not produce a more human-like or consistent profile in either direction, but the two Qwen sizes agree with each other on both illusions where any within-Qwen agreement exists (graded on Müller-Lyer, erratic on Ebbinghaus) while LLaVA diverges on both (erratic, graded respectively) – consistent with, though far from conclusive evidence for, the encoder rather than decoder scale being the more relevant locus of a model's illusion-response *shape*. This does not resolve the locus of *magnitude* (the fitted PSE) at all; both questions need a design varying encoder and decoder independently at larger N than the three checkpoints here (Section 8).

8 LIMITATIONS

Our study is limited to four classical illusions with a single scalar manipulated variable each; illusions with multiple interacting variables (e.g., Zöllner, Poggendorff) may require a different formalism. Human reference data come from lab conditions that differ from a VLM's image-token pipeline, though our qualitative monotonicity/step-function distinction is less sensitive to this than absolute offsets are. VLM responses are also sensitive to prompt phrasing and resolution beyond the label/position bias we counterbalance: two bounded $R = 8$ subset checks (Appendix F) each flipped one model's classification, a different model each time – evidence against over-interpreting any single wording/resolution rather than that one model is more robust. Our model set covers three open-weight checkpoints across two vision-encoder families, short of the closed-model coverage a fully general claim about "VLMs" would need. Most importantly, neither the forced-choice classifications nor confidence-coupling are stable to resampling even after doubling R: across our three reruns, only 1 of 8 comparable forced-choice cells and no confidence-coupling pattern at all keeps the same classification throughout (Appendix F). Treat every per-cell classification in Table 1 as illustrative of a broader pattern, not an individually reproducible measurement.

9 CONCLUSION

We introduced ILLUSIONCURVE-BENCH, a parametric extension of classical-illusion evaluation replacing single counterfactual images with dense stimulus-magnitude sweeps, separating two failure modes binary-accuracy benchmarks conflate: graded, human-like bias and categorical, memorization-driven pattern matching. Applied to three open-weight VLMs across two vision-encoder families with label/position confounds counterbalanced out, binary-accuracy's confident verdicts mask response functions that are overwhelmingly non-human-like (9 of 12 fits erratic or unfittable) and confidence signals essentially decoupled from ambiguity – doubling the repeat count sharpened rather than resolved this, changing which cells looked dip-like for a third time. Binary illusion accuracy is compatible with, and in our results is generated by, response functions resembling neither graded human perception nor a stable, single-pass measurement. We release our sweep generators, stimuli, model responses, and analysis code.

Reproducibility Statement

We release code for stimulus generation (a from-scratch parametric renderer, since the original VLMBias generator of Vo et al. (2025) was not available to us), evaluation prompts, raw model responses, and psychometric curve-fitting scripts at [MASKED]. All three models are pinned open-weight Hugging Face Hub checkpoints (`Qwen/Qwen2.5-VL-3B-Instruct`, `Qwen/Qwen2.5-VL-7B-Instruct`, `llava-hf/llava-v1.5-7b-hf`), which do not drift the way a closed-model API can; the exact revision hash resolved by each is logged alongside the raw responses at run time. Closed-model results, once added, will need their exact API version strings and query dates logged for the same reason.

Ethics Statement

This work evaluates existing vision-language models on synthetically generated geometric stimuli and does not involve human subjects, personal data, or dual-use capabilities. All compute is local, open-weight GPU inference on hardware already available to the authors (Section 5); no third-party model API was queried and no associated data-sharing or usage-policy considerations arise for this submission. Human reference values (Appendix D) are secondary citations to previously published psychophysics studies, not new data collection from human subjects.

AI Use Statement

An LLM-based coding agent was used extensively and substantively throughout this project: implementing the generators/harness/fitting/analysis pipeline and validation scripts released with the paper, executing the local GPU inference runs and statistical analyses (including Appendix E's bootstrap check), and drafting and revising the text, including material added in response to review feedback. The human author directed the research questions, made all methodological decisions the agent surfaced as options, reviewed the generated code and text, and is responsible for the accuracy of all claims here – including having independently spot-checked Section 6's numeric claims against the underlying data. No LLM was used as an author, and no LLM-generated content was included without human review.

References

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations (ICLR)*, 2019.

George A. Gescheider. *Psychophysics: The Fundamentals*. Lawrence Erlbaum Associates, 3rd edition, 1997.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Hester Knol, Raoul Huys, Jean-Christophe Sarrazin, and Viktor K. Jirsa. Quantifying the Ebbinghaus figure effect: target size, context size, and target-context distance determine the presence and direction of the illusion. *Frontiers in Psychology*, 6:1679, 2015. doi: 10.3389/fpsyg.2015.01679.

Klaus Landwehr. Deconfounded and mixed-symmetry versions of the Ponzo illusion figure. *Vision Research*, 202:108143, 2023. doi: 10.1016/j.visres.2022.108143. 33/120 PSE-fitting failure rate corroborated at abstract level in this session; full text was paywalled, not independently read.

Na Lin, Bichi Chen, Maoyuan Yang, Fan Lu, and Ruzhi Deng. Low vision aids and age are associated with Müller-Lyer illusion in congenital visually impaired children. *Frontiers in Psychology*, 14:1278554, 2023. doi: 10.3389/fpsyg.2023.1278554.

Kyriaki Mikellidou and Peter Thompson. The vertical-horizontal illusion: Assessing the contributions of anisotropy, abutting, and crossing to the misperception of simple line stimuli. *Journal of Vision*, 13(8):7, 2013. doi: 10.1167/13.8.7.

Frank Restle and Jozef Decker. Size of the Mueller-Lyer illusion as a function of its dimensions: Theory and data. *Perception & Psychophysics*, 21(6):489–503, 1977. doi: 10.3758/BF03198729.

Haz Sameen Shahgir, Khondker Salman Sayeed, Abhik Bhattacharjee, Wasi Uddin Ahmad, Yue Dong, and Rifat Shahriyar. Illusionvqa: A challenging optical illusion dataset for vision language models, 2024.

An Vo, Khai-Nguyen Nguyen, Mohammad Reza Taesiri, Vy Tuong Dang, Anh Totti Nguyen, and Daeyoung Kim. Vision language models are biased, 2025. No peer-reviewed venue confirmed as of the most recent revision (April 2026) at the time of writing; verify before camera-ready in case a conference version appears.

A Appendix: Full Prompt Templates

Forced-choice perceptual judgment (noun/property substituted per illusion family – "circle"/"larger" for Ebbinghaus, "line segment"/"longer" for Müller-Lyer and Ponzo, "segment"/"longer" for Vertical–Horizontal):

```
The image shows two [circle/line segment]s, labeled A
and B. Which [circle/line segment] is [larger/longer],
A or B? If they appear the same size, answer Equal.
Respond with exactly one word:  A, B, or Equal.
```

Same/different identification (identical across all four illusion families; the reference image is always the $x = 0$ veridical stimulus for that illusion):

```
You will be shown two images:  a reference image,
followed by a test image.  Is the test image
perceptually the same as the reference image, or
different?  Respond with exactly one word:  Same or
Different.
```

Reworded forced-choice variant used only for the prompt-phrasing robustness check (Section 8), not the primary protocol – same neutral semantic content, opposite A/B mention order:

```
Compare the two [circle/line segment]s shown in this
image, marked A and B. Is B or A the [larger/longer]
one?  If they look equal in size, respond Equal.
Reply with a single word only:  A, B, or Equal.
```

B Appendix: Generator and Sweep Parameters

All four generators expose a single scalar sweep variable $x \in [-1, 1]$, swept at $K = 11$ evenly spaced steps (including the exact veridical point $x = 0$), implemented from scratch in `generators.py` (no VLMBias generator code was available to extend, per Section 3.2's correction). Sign convention: $x > 0$ biases the forced-choice judgment toward "B"; $x < 0$ toward "A"; $x = 0$ is veridical/neutral. Every generator additionally renders a `swap_labels` variant that redraws the opposite text at each of the two fixed slots without moving any geometric parameter (the "Label-assignment and positional counterbalancing" paragraph in Section 5), used to counterbalance label-identity response bias out of the fitted PSE; Ponzo and VHI additionally render a `pos_swap` variant that instead moves the swept-length element to the opposite physical slot without changing any label, counterbalancing positional bias (not meaningful for Ebbinghaus/Müller-Lyer – see that same paragraph). Table 3 gives the exact per-illusion parametrization and which human reference study (Appendix D) each sweep's geometry was designed to be comparable with. Ponzo sweeps physical length ratio rather than position (Section 3.2) because the near/far positions are what makes the perspective illusion-inducing context present at every step; sweeping position itself would force the two segments to coincide exactly at the symmetric/veridical point, which cannot be rendered as two distinct labeled elements.

Curve-fitting solver settings (Section 4.1). `scipy.optimize.minimize` with the Nelder-Mead method, `xatol` = `fatol` = 10^{-6}, up to 5000 iterations. Unfittable if fewer than 4 sweep steps have at least one valid (non-Equal, non-invalid) response, the optimizer's own convergence check fails, or it converges to α more than one full sweep-width outside $[-1, 1]$. Monotonicity violation threshold: 30% of adjacent-step pairs. Step-like threshold: fitted $|\beta|$ below 5% of the sweep's x-range.

C Appendix: Per-Model Full Psychometric Curves

Figure 1 in the main text already gives the full per-model, per-illusion psychometric grid at native resolution (three models $\times$ all four illusions, 12 panels total); with only three models in this pilot

Table 3: Exact sweep parametrization for each illusion family. $x \in [-1, 1]$ throughout; slot 1/slot 2 refer to the fixed rendering slot (not the on-screen text, which `swap_labels` may exchange between slots, nor – for Ponzo/VHI – the physical position, which `pos_swap` may exchange independently; the "Label-assignment and positional counterbalancing" paragraph in Section 5).

Illusion	Swept geometry	Fixed context	Human ref. study
Ebbinghaus	Central circle radius fixed; slot 1 context-circle radius $= r_0(1+0.8x)$, slot 2 $= r_0(1-0.8x)$, r_0 = context radius at $x = 0$ (100% of target size, matching Knol et al. (2015)'s own no-illusion baseline; sweep spans their tested 20%-180% context-ratio range exactly)	7 context circles/side, orbit radius $3.08\times$ central radius (matching Knol et al. (2015)'s minimum-distance formula at their largest tested context ratio)	Knol et al. (2015)
Müller-Lyer	Fin angle from inward direction: $\theta_1(x) = 90° - 70°x$, $\theta_2(x) = 90° + 70°x$	Shaft length fixed; fin length $= 0.11\times$ image size	Lin et al. (2023); sweep range informed by Restle & Decker (1977)
Ponzo	Slot 1 length $= L_0(1 + 0.35x)$ (swept); slot 2 length fixed at L_0. `pos_swap` moves slot 1 (swept) from the apex ("far") position to the base ("near") position and vice versa, independent of label	Fixed converging-line apex and near/far target positions (Section 3.2)	Landwehr (2023) (geometry not matched exactly – Section 3.2 deviation)
Vertical–Horizontal	Slot 1 length $= L_0(1 + 0.4x)$ (swept); slot 2 length fixed at L_0. `pos_swap` makes the horizontal segment the swept one and the vertical the fixed reference (default: swept=vertical), independent of label	Slot 2 (fixed-reference) segment orthogonal to slot 1	Mikellidou & Thompson (2013), inverted-T configuration

there is no additional per-page detail to add beyond that figure. The full confidence-curve grid (three models $\times$ all four illusions, 12 panels, of which Figure 3 shows two representative panels) is released alongside the code as `paper/figures/confidence_<model>_<illusion>.pdf`.

D Appendix: Human Reference Data Sources

Table 4 summarizes the human psychophysical reference values used for comparison in Section 5 and Table 1. These were re-verified directly against primary sources (not search-engine snippets) in this revision; see `human_refs.py` for the per-entry verification notes. Two corrections from an earlier draft of this table are worth flagging explicitly, since they change the citations rather than just the numbers: the Ebbinghaus 0–10% figure is correctly attributed to Knol et al. (2015) (an earlier draft cited a non-existent "Doyon 2015"), and the Vertical–Horizontal cross-configuration figure (7% underestimation, PSE 5.65°) is correctly attributed to Mikellidou & Thompson (2013) rather than to Prinzmetal & Gettleman (1993), whose actual reported numbers concern monocular-vs-binocular viewing ($\approx$2.47%/$-$1.36%) and do not match the cross-configuration claim. Note that none of these illusions has a single universal magnitude; each reported value is tied to a specific

stimulus configuration (fin angle, context-to-target size ratio, line orientation, viewing conditions), so ILLUSIONCURVE-BENCH's own sweep geometry for each illusion should be matched as closely as possible to the configuration under which the corresponding human reference value was obtained, and any deviation should be reported explicitly alongside the comparison in Table 1 – most notably Ponzo's length-ratio-vs-position deviation (Section 3.2).

E APPENDIX: BOOTSTRAP CLASSIFICATION STABILITY

Section 6.1 shows that the psychometric classification of a given (model, illusion) cell changes across independent reruns of the protocol. That comparison cannot by itself distinguish two explanations: ordinary finite-sample noise at $R = 16$ (the true response process is unchanged, but any given $R = 16$ draw is one noisy sample among many plausible ones), versus some deeper non-stationarity (the process itself differs between independent inference sessions). `bootstrap_classification_stability.py` (released with the code) addresses this using only the data already collected, with no new model queries: for each cell, it resamples the observed per-step response-category counts (a multinomial bootstrap over the actual A/B/Equal/invalid counts at each of the 11 steps, preserving each step's total n) 1000 times, refits the psychometric function on each resample, and reports what fraction reproduce the classification actually observed at $R = 16$. This asks a narrower question than the three-rerun comparison: holding the true process fixed, how often would *this exact* classification arise by chance at this sample size?

Table 5 gives the results for all 12 cells. The pattern is stark and, we think, informative about which of the paper's findings to trust most. All three graded/human-like classifications are the least bootstrap-stable in the entire dataset ($P(\text{same class}) = 0.20$–$0.40$): resampling the same underlying $R = 16$ data, with nothing about the model or protocol changed, reclassifies these cells as erratic or unfittable roughly 60–80% of the time. The unfittable classifications are also comparatively fragile (0.38–0.44), most often flipping to erratic under resampling. The erratic/non-monotonic classifications are, by contrast, the most bootstrap-robust (0.52–0.85, and above 0.6 for three of the four Qwen2.5-VL-7B cells specifically). This gives a quantitative rather than purely qualitative basis for weighting the paper's claims: the central finding that most model–illusion response curves are erratic and uncorrelated with x rests on the most bootstrap-stable part of the data, while the handful of graded/human-like cells we report in Table 1 are, by this same measure, plausibly artifacts of the specific $R = 16$ draw rather than a reproducible property of the underlying model. We regard this as a more precise version of the caution already raised in Section 6.1 and Section 8, not a contradiction of it: the instability observed across independent reruns is at least partly explained by exactly the kind of finite-sample noise this bootstrap quantifies, and it is concentrated in precisely the cells (graded, unfittable) that a reader should already be treating with the most skepticism.

F APPENDIX: ADDITIONAL ROBUSTNESS AND CROSS-RUN STABILITY DETAILS

This appendix collects the figures and detailed numbers referenced from the main text but moved here to keep the main body concise: example stimuli, the confidence dip-vs-collapse comparison figure, the full three-run comparison for both the forced-choice and confidence signals (summarized in Section 6.1/6.2), the prompt-phrasing/resolution subset checks (summarized in Section 8), and the positional-counterbalancing validation numbers (summarized in Section 5).

Three-run forced-choice classification comparison. Table 6 compares the classification of each (Qwen-3B/Qwen-7B $\times$ illusion) cell across an initial uncounterbalanced $R = 8$ pilot, a counterbalanced $R = 8$ rerun, and the counterbalanced $R = 16$ run reported in the main text. Only Vertical–Horizontal/Qwen-7B (erratic throughout) keeps the same classification in all three.

Three-run confidence-coupling comparison. The uncounterbalanced $R = 8$ run showed Qwen-7B with a same/different rate that dipped near $x = 0$ on Ebbinghaus and Müller-Lyer specifically, collapsing to a near-constant "Same" response on the other two illusions. The counterbalanced $R = 8$ rerun instead showed Qwen-7B collapsed on all four illusions, with the only dip-like pattern in that run appearing on LLaVA/VHI. The counterbalanced $R = 16$ run (main text, Section 6.2)

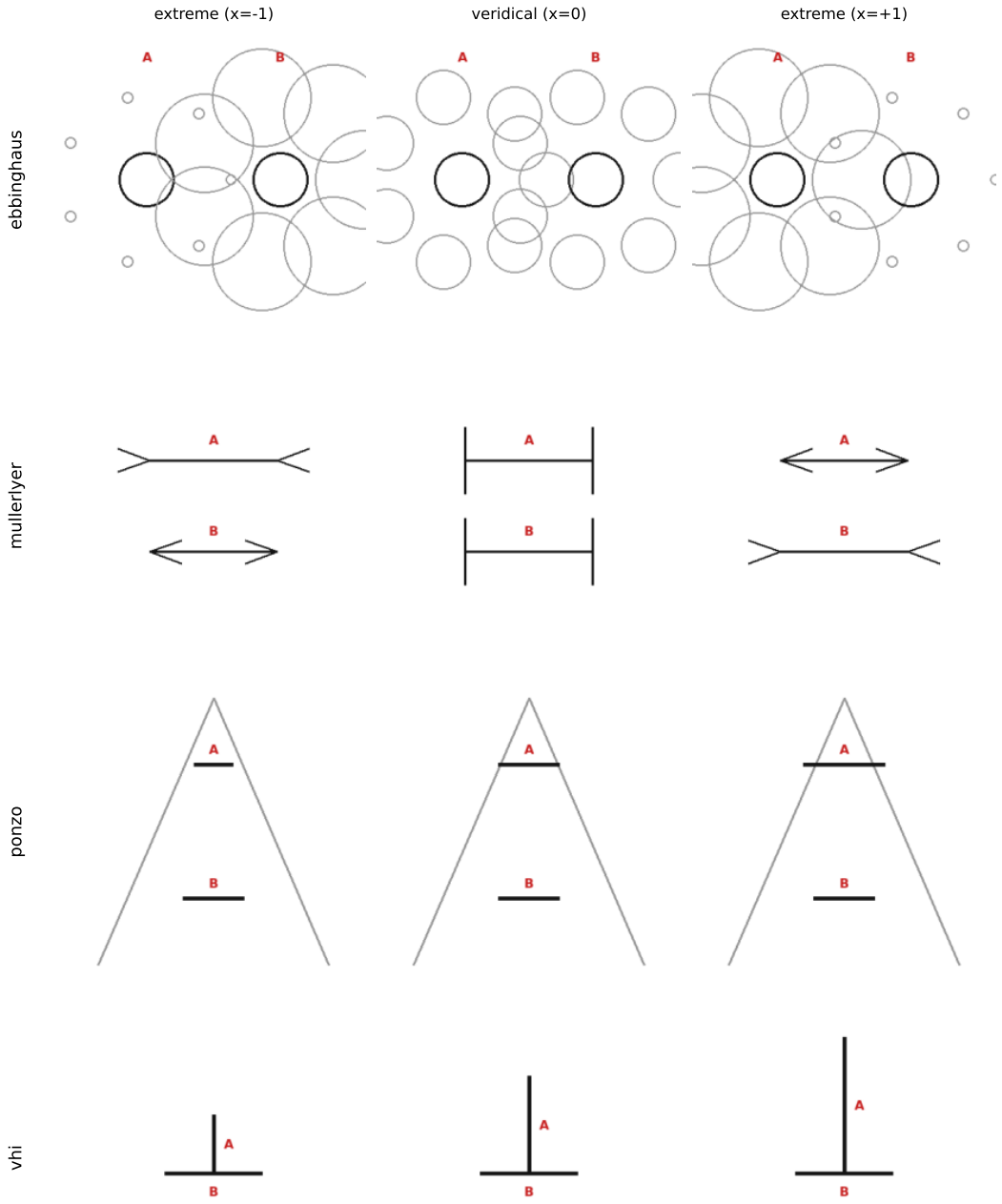


Figure 2: Example stimuli from three sweep steps (extreme, veridical, extreme) for each of the four illusion families in ILLUSIONCURVE-BENCH.

shows a third pattern again: Qwen-3B newly dipping on Ebbinghaus, LLaVA still dipping on VHI, and Qwen-7B collapsed everywhere except a marginal, floor-level dip on Ebbinghaus. LLaVA/VHI is the only cell that dipped in more than one of the three runs.

Confidence dip examples, full numbers (Section 6.2). Qwen2.5-VL-3B on Ebbinghaus: 0.5625 at $x = -1$, 0.75 at $x = -0.8$, 0.1875 at $x = 0$ (the sweep's minimum), 0.625 at $x = +0.2$, 0.5625

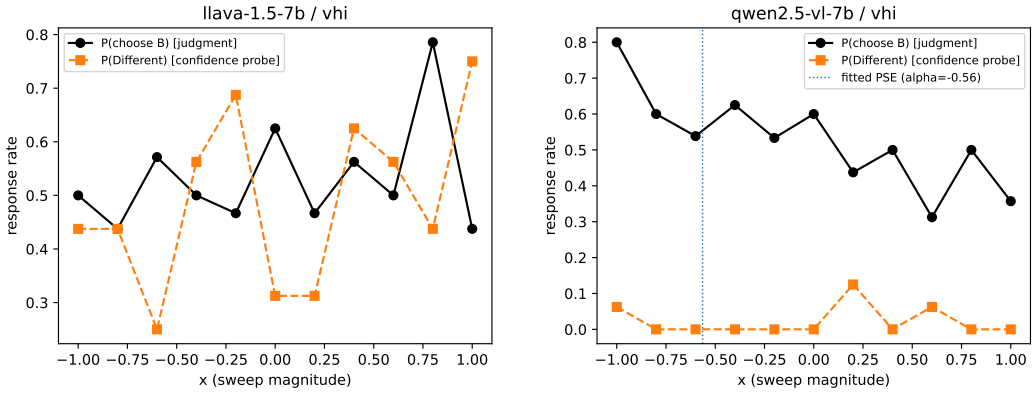


Figure 3: Perceptual judgment curve versus same/different identification rate as a function of stimulus magnitude, both on Vertical–Horizontal: LLaVA-1.5-7B (left, one of the two cells across all 12 model–illusion pairs coming closest to a confidence dip near $x = 0$, though its true minimum sits off-center at $x = -0.6$ – Section 6.2) versus Qwen2.5-VL-7B (right, collapsed/magnitude-insensitive identification response, representative of that model's pattern on all four illusions). Full grid for all three models and all four illusions is released alongside the code.

at $x = +1$ – both endpoints coincidentally equal, but the intervening steps do not monotonically approach the $x = 0$ minimum. LLaVA-1.5-7B on Vertical–Horizontal: 0.4375 at $x = -1$, 0.3125 at $x = 0$, 0.75 at $x = +1$ – below both endpoints at $x = 0$, but the curve's true minimum (0.25) occurs off-center at $x = -0.6$. Qwen2.5-VL-7B on Ebbinghaus: $0.0625/0.0/0.0625$ at the same three points, the same directional pattern at floor-level magnitude. For contrast, Qwen2.5-VL-7B on Ponzo *peaks* at $x = 0$ rather than dipping: $0.0625/0.125/0.0$ at the two extremes and center.

Prompt-phrasing robustness check. Rewording the Müller-Lyer forced-choice prompt (Appendix A gives the exact wording, reversing which of A/B is mentioned first) and rerunning at $R = 8$ (versus $R = 16$ for the primary result – not a matched comparison) gave: Qwen2.5-VL-3B $\alpha = +40.0\%$, erratic (primary: graded, -144.5%); Qwen2.5-VL-7B $\alpha = -114.9\%$, graded, 66% Equal-rate (primary: graded, -181.7%, 86% Equal-rate); LLaVA-1.5-7B $\alpha = +169.3\%$, erratic (primary: erratic, $+291.8\%$).

Resolution robustness check. Rendering Ebbinghaus at 384px instead of 768px and rerunning at $R = 8$ gave: Qwen2.5-VL-3B $\alpha = +15.4\%$, erratic, 41% Equal-rate (primary: erratic, -18.6%, 30% Equal-rate); Qwen2.5-VL-7B $\alpha = -1.5\%$, erratic, 0% Equal-rate (primary: erratic, -0.0%, 0% Equal-rate); LLaVA-1.5-7B $\alpha = -65.1\%$, erratic, 14% Equal-rate (primary: graded, -176.8%, 5% Equal-rate).

Positional-counterbalancing validation. `validate_position_counterbalance.py` simulates a synthetic model with a known true PSE of 0.15 and a 60% constant positional bias (always favoring whichever slot sits at the Ponzo apex, regardless of which role is there). The uncounterbalanced fit recovers $\alpha = -0.63$ (error 0.78, more than $5\times$ the true PSE's magnitude); the `pos_swap`-counterbalanced fit recovers $\alpha = 0.13$ (error 0.02), the same order-of-magnitude correction as the label-bias validation in Section 5.

Transient-OOM retry rationale (Section 5). This box is shared with other users' jobs on all 4 GPUs; their memory usage can grow after our model has already loaded, mid-generation, colliding with ours. `hf_query` retries up to 3 times (clearing the CUDA cache between attempts) on a transient out-of-memory error rather than letting one spike kill a multi-hour sweep's progress – confirmed necessary in practice: the Ebbinghaus re-run for this submission hit exactly this failure mode twice before succeeding on retry.

Table 4: Human psychophysical reference values for the four illusion families in ILLUSIONCURVE-BENCH. All magnitudes are reported as illusion strength (percent deviation from the veridical point of equality) unless otherwise noted.

Illusion	Reference	Method	Reported value / notes
Ebbinghaus	Knol et al. (2015)	Staircase + method of constant stimuli	Illusion magnitude ranges from ∼0% to (exceptionally) 10% of target size; PSE consistent across staircase and constant-stimuli methods. Magnitude is strongly dependent on context-to-target size ratio and target-context distance: their 3×3×3+3 factorial design tested context sizes at 20%/100%/180% of target size and, at each target size, a minimum target-context distance of 1.1×(target radius + largest context radius), with the largest illusion magnitudes reported at that smallest distance. `generators.render_ebbinghaus` now sweeps context ratio over exactly the tested 20%-180% range with orbit radius matching their minimum-distance formula (Table 3), resolving this configuration match, though our two-sided forced-choice sweep still differs from their single-figure-vs-reference paradigm.
Müller-Lyer	Lin et al. (2023)	Two-alternative forced choice	Illusion intensity ≈5.42% (SD 3.81) in sighted controls, standard line 250px, arrowhead 20% of line length (corrected from 25% in an earlier draft). Restle & Decker (1977) give magnitude as a function of fin angle/length for justifying sweep range design.
Ponzo	Landwehr (2023)	2AFC longer/shorter judgment, logistic regression PSE fit	PSE fitted at the 50% point of the psychometric function; exact magnitude not independently confirmed (source full text paywalled – abstract-level corroboration only). PSE fitting failed in 33/120 cases in the source study – an unfittable-curve base rate worth citing in our own curve-fitting failure analysis (Section 6.1).
Vertical–Horizontal	Mikellidou & Thompson (2013)	Method of constant stimuli / 2AFC	Inverted-T configuration: ∼9% overestimation of the vertical segment (PSE 6.62° vs. veridical 6.1°, $p < 0.01$). Cross (+) configuration: ∼7% underestimation (PSE 5.65°) – both configurations are reported in this same source. Broader literature reports a 5–10% range for the standard planar illusion. We use the inverted-T configuration as our primary reference since it matches this project's generator geometry (`generators.render_vhi`).

Table 5: Bootstrap classification stability for each of the 12 (model, illusion) cells at $R = 16$: the fraction of 1000 multinomial-resampled refits of that cell's own data that reproduce its actual classification, and the most common alternative classification when they don't.

Model	Illusion	Actual classification	P(same class)	Top alternative
qwen2.5-vl-3b	ebbinghaus	erratic/non-monotonic	0.85	graded/human-like (0.11)
qwen2.5-vl-3b	mullerlyer	graded/human-like	0.20	erratic/non-monotonic (0.53)
qwen2.5-vl-3b	ponzo	unfittable	0.44	erratic/non-monotonic (0.47)
qwen2.5-vl-3b	vhi	unfittable	0.39	erratic/non-monotonic (0.53)
qwen2.5-vl-7b	ebbinghaus	erratic/non-monotonic	0.74	step-like/categorical (0.23)
qwen2.5-vl-7b	mullerlyer	graded/human-like	0.40	unfittable (0.34)
qwen2.5-vl-7b	ponzo	erratic/non-monotonic	0.81	graded/human-like (0.12)
qwen2.5-vl-7b	vhi	erratic/non-monotonic	0.65	graded/human-like (0.34)
llava-1.5-7b	ebbinghaus	graded/human-like	0.21	erratic/non-monotonic (0.53)
llava-1.5-7b	mullerlyer	erratic/non-monotonic	0.52	unfittable (0.37)
llava-1.5-7b	ponzo	unfittable	0.38	erratic/non-monotonic (0.52)
llava-1.5-7b	vhi	unfittable	0.38	erratic/non-monotonic (0.53)

Table 6: Forced-choice classification across all three protocol runs, Qwen-3B/Qwen-7B only (LLaVA was not present in the first two runs).

Illusion	Model	Uncounterbal. $R = 8$	Counterbal. $R = 8$	Counterbal. $R = 16$
ebbinghaus	3B	graded (-25.2%)	step-like ($+15.6\%$)	erratic (-18.6%)
ebbinghaus	7B	erratic (-20.0%)	step-like ($+11.1\%$)	erratic (-0.0%)
mullerlyer	3B	erratic (-122.6%)	graded (-7.4%)	graded (-144.5%)
mullerlyer	7B	step-like ($+88.3\%$)	graded (-65.4%)	graded (-181.7%)
ponzo	3B	erratic (-108.0%)	erratic (-28.2%)	unfittable
ponzo	7B	unfittable	erratic ($+68.1\%$)	erratic (-87.7%)
vhi	3B	step-like ($+62.4\%$)	erratic ($+113.9\%$)	unfittable
vhi	7B	erratic (-133.8%)	erratic ($+87.0\%$)	erratic (-56.4%)