DO GENERATIVE PRIORS ALIGN WITH HUMAN NATURALNESS PERCEPTION?

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual generative models learn probability distributions over natural images, but whether their native priors directly capture human perception of image naturalness remains unresolved. Here, we probe these priors through native prediction errors across 25 open image and video generators. Because raw single-image losses are dominated by scene content and visual complexity, we evaluate directional loss differences using content-preserving, paired relational interventions that selectively disrupt facial configurations or physical illumination consistency while limiting changes in low-level image statistics. In representative conditions, these loss differences reproduce selective human sensitivities and tolerances, capturing the classic Thatcher effect on faces and shape-dependent responses to illumination inconsistencies. Crucially, they reliably track continuous human naturalness judgments across individual stimulus pairs (peaking at $r = .84$ on faces and $.64$ on physical scenes). Across the evaluated models, overall sensitivity to these violations broadly covaries with human alignment, yet the two dissociate along denoising schedules, with fine-grained perceptual alignment typically peaking before sensitivity. Importantly, generative loss differences retain positive partial correlations with human judgments after controlling for feature distances from frozen vision encoders and standard image quality metrics. Together, these findings demonstrate that learning visual distributions yields generative loss landscapes that capture distinct aspects of human naturalness perception.

1 INTRODUCTION

Human perception of image naturalness draws heavily on sensitivity to the statistical regularities of natural visual environments: cognitive theories formalize this as inference against internal distributional priors acquired through visual experience (Yuille & Kersten, 2006). Just as human vision relies on approximate statistical expectations rather than exact physical simulation, visual generative models produce strikingly realistic images from learned image distributions while frequently violating strict physical and geometric constraints (Farid, 2022; Sarkar et al., 2024). This parallel raises the possibility that generative priors, learned from visual distributions rather than explicit physical simulation, naturally capture the selective sensitivities and tolerances of human vision.

Prior studies utilizing generative models for perceptual assessment have largely treated them as feature extractors, training supervised regressors on internal activations (Fu et al., 2024; Raviv & Chechik, 2024) or engineering tuning-free attention-based metrics (Song et al., 2025). To test whether generative models naturally internalize human-aligned perceptual regularities without relying on such task-specific readouts, we turn directly to their native prediction objectives. In diffusion and flow architectures, denoising and velocity losses relate to score matching and variational bounds on data log-likelihood (Ho et al., 2020; Kingma & Gao, 2023; Esser et al., 2024), providing an empirical surrogate for negative log-likelihood that reflects an image's plausibility under the learned distribution (Clark & Jaini, 2023). We therefore evaluate whether these native prediction errors directly track human naturalness ratings in a zero-shot manner, probing the perceptual sensitivities captured by native generative prediction objectives.

Evaluating native loss directly, however, faces a known obstacle: high-dimensional density estimates and unconditioned prediction errors are sensitive to low-level image complexity, background statistics, and data geometry, which can decouple density from perceived plausibility (Nalisnick et al., 2019;

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.

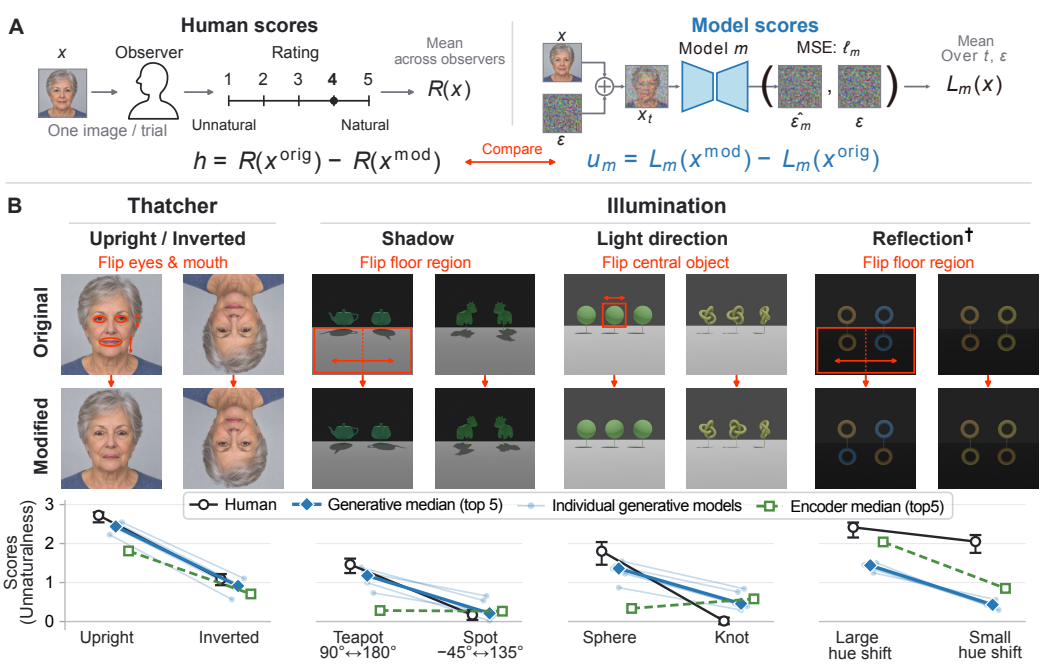

Figure 1: **Overview of experimental paradigm and paired response profiles.** (A) Measurement pipelines. Left: human naturalness ratings are averaged across observers to define the paired perceptual effect h. Right: native prediction errors from frozen models are averaged over matched timesteps and noise to obtain image losses and the paired directional loss difference u_m (schematic illustrates a noise-predicting diffusion model; prediction targets vary across architectures). (B) Controlled stimulus pairs and representative response profiles. Content-preserving relational modifications selectively flip local elements (eyes/mouth, floor regions, or central objects; red boxes), leaving overall scene composition intact. Left (Thatcher task): both humans and generative models penalize facial distortions far more strongly in upright than in inverted faces. Right (Illumination task): selected factor-level contrasts across cast shadows, light directions, and surface reflections; generative models mirror human sensitivities to shadows and lighting. †Reflection color is included as a representative divergence, where humans remain sensitive to small hue mismatches while generative responses decrease. Markers indicate human means with 95% bootstrap intervals (Appendix E), top-five generative models (median and individuals), and top-five vision encoders (median). Complete profiles appear in Appendix J. Responses are standardized by task-wide standard deviation. Face images depict a synthetic person solely for illustrative purposes and were not used in experiments.

Serrà et al., 2020; Ren et al., 2019; Kamkari et al., 2024). Consistent with these concerns, in our control evaluation on a standard distortion benchmark (CSIQ (Larson & Chandler, 2010)), native prediction errors fail to track human quality judgments across distortion types, and removing visual elements like cast shadows can lower loss even though the scene becomes physically unnatural (Appendix F). To separate relational plausibility from low-level statistical shifts, we adopt psychophysical paradigms based on content-preserving relational interventions (Thompson, 1980; Ostrovsky et al., 2005; Nightingale et al., 2019). Instead of adding or deleting objects, our manipulations alter spatial or physical configurations by reflecting existing image elements, which leaves coarse image statistics largely intact. The paired directional loss difference ($L_{\text{modified}} - L_{\text{original}}$) then controls for shared scene composition, focusing evaluation on whether the model penalizes the relational violations that reduce human-perceived naturalness.

We evaluate this approach across two controlled domains: the Thatcher face task (feature inversions in upright and inverted faces) and the Illumination task (shadow, reflection, and light-direction interventions in rendered scenes; Figure 1). In representative response profiles, paired generative loss differences capture characteristic human perceptual patterns—such as the classic Thatcher effect on faces and shape-dependent shadow sensitivities—that are not exhibited by standard vision encoders, while also revealing specific divergences (Figure 1B). To test whether these correspondences reflect

systematic perceptual alignment, we evaluate 25 open image and video generators on individual stimulus pairs, comparing native loss differences against human naturalness ratings, 12 frozen vision and multimodal encoders, and ten standard image quality metrics.

Our investigation yields three primary findings. First, zero-shot directional loss differences reliably track continuous human ratings across both tasks (peaking at $r = .84$ on faces and $.64$ on physical scenes) and concentrate spatially within manipulated scene regions, whereas unconditioned raw losses fail to do so. Second, overall violation sensitivity broadly covaries with human alignment across the evaluated models, yet the two decouple along denoising schedules, with perceptual alignment typically peaking earlier than sensitivity. Moreover, this loss-based alignment broadly corresponds to public generation benchmarks after controlling for sensitivity. Third, controlling for representations from frozen vision encoders and standard quality metrics confirms that native generative losses retain non-redundant, human-aligned perceptual structure beyond that captured by these baseline readouts. Taken together, our work provides empirical evidence for a functional convergence between visual generative priors and human perception, showing that modern generative loss landscapes capture key aspects of the selective sensitivities and tolerances of human naturalness judgments.

2 Related Work

2.1 Generative Models as Perceptual Priors

Evaluating departures from natural image statistics is a long-standing approach in perceptual assessment. Classical no-reference image quality assessment (NR-IQA) measures distortions as deviations from parametric models of natural scene statistics (Mittal et al., 2012; 2013; Saad et al., 2012). From a Bayesian perspective, visual quality assessment formalizes this principle as statistical inference against learned priors of natural visual environments (Duanmu et al., 2021), echoing cognitive theories of biological vision as Bayesian inference or inverse graphics (Yuille & Kersten, 2006; Yildirim et al., 2020). Modern deep generative models scale this principle by fitting high-dimensional distributions directly on large image corpora. However, raw density estimates and unconditioned prediction errors do not directly track perceptual quality. Likelihood-based deep models frequently assign higher likelihoods to out-of-distribution inputs than to in-distribution training data (Nalisnick et al., 2019; Kirichenko et al., 2020), reflecting sensitivity to input complexity (Serrà et al., 2020), background statistics (Ren et al., 2019), and manifold geometry (Kamkari et al., 2024).

To bypass these density confounds, recent approaches treat generative models as feature extractors, using supervised regressors on internal activations (Fu et al., 2024; Raviv & Chechik, 2024) or tuning-free attention correspondences (Song et al., 2025). While effective for metric engineering, these feature-derived methods do not evaluate whether the model's learned distribution itself aligns with human naturalness perception. Rather than engineering an applied IQA metric, our goal is a scientific inquiry: testing whether human perceptual priors are reflected directly within the native loss landscapes of generative models, using content-preserving relational interventions to probe prediction errors without intermediate readouts.

2.2 Physical World Models vs. Perceptual Naturalness

Several recent benchmarks evaluate whether generative models learn physical and causal regularities as predictive world models. While VideoPhy evaluates physical commonsense in video generation (Bansal et al., 2025) and Physics-IQ tests real-world physical continuation (Motamed et al., 2026), LikePhys (Yuan et al., 2026) and YoCausal (Xie et al., 2026) use paired denoising-loss contrasts to test model preferences for physically valid or temporally plausible sequences, verifying their benchmarks against aggregate human model rankings. However, adherence to objective physical laws does not directly imply alignment with human perceptual naturalness. Psychophysical studies demonstrate that human observers frequently fail to detect or tolerate substantial physical inconsistencies in lighting, shadows, and reflections (Mamassian, 2004; Ostrovsky et al., 2005; Cavanagh, 2005; Wilder et al., 2019; Nightingale et al., 2019). A model that strictly penalizes all physical departures may therefore diverge from human naturalness judgments. Rather than evaluating whether models act as strict physical simulators, our Illumination task examines whether generative loss landscapes mirror the selective tolerances that characterize human perception, testing item-level correspondence with continuous human naturalness ratings under controlled physical interventions.

2.3 Visual Illusions in Deep Networks

Visual illusions and perceptual biases provide a classic benchmark for testing whether artificial neural networks reproduce biological visual phenomena (G'omez-Villa et al., 2020; Jacob et al., 2021; Shahgir et al., 2024). Prior computational studies have investigated these effects across several paradigms: convolutional filters adapt to natural statistics to reproduce low-level brightness illusions (G'omez-Villa et al., 2020), networks trained on edge-based image reconstruction replicate surface filling-in illusions (Saha et al., 2026), and diffusion models reveal human-like brightness shifts along their DDIM inversion trajectories (Gomez-Villa et al., 2025). In higher-level perception, configural phenomena such as the Thatcher illusion have been investigated in feedforward neural networks (Jacob et al., 2021), while structured inverse graphics models have captured other face-perception illusions such as the hollow-face effect (Yildirim et al., 2020). In addition, multimodal benchmarks probe illusions via linguistic question answering (Shahgir et al., 2024). Extending this line of inquiry, we test whether characteristic human perceptual biases are expressed directly within the native loss landscapes of visual generative models.

3 Methods

3.1 Stimulus Design and Psychophysical Experiments

Probing human naturalness perception across two domains. To investigate whether generative priors reflect human naturalness perception, we examine two distinct visual domains: face configurations and physical scene illumination. In face perception, observers exhibit pronounced orientation dependence: in the Thatcher illusion, inverted eyes and mouth appear grotesque in an upright face, but this distortion becomes difficult to detect when the entire face is inverted (Thompson, 1980; de Haas & Schwarzkopf, 2018). Testing Thatcherized faces evaluates whether generative models detect these localized feature inversions and whether their sensitivity exhibits orientation dependence. In physical scenes, observers can fail to detect substantial inconsistencies in illumination, shadows, and reflections (Mamassian, 2004; Ostrovsky et al., 2005; Cavanagh, 2005; Wilder et al., 2019; Nightingale et al., 2019). Testing illumination interventions determines whether models identify physical inconsistencies at all, and whether their responses reflect physical mechanics strictly or follow human perceptual tolerances and naturalness ratings. Evaluating models across both domains tests both basic violation sensitivity and its correspondence to human naturalness judgments.

Content-preserving relational interventions. To probe relational regularities while limiting low-level statistical shifts, we design two tasks using *content-preserving relational interventions*. Existing image content is rearranged while depicted objects remain in place, avoiding object deletion or insertion. This design largely preserves color distributions and edge content, reducing the density confounds observed in evaluation of raw native losses. The Thatcher task contains 140 public-domain face scenes from FFHQ (Karras et al., 2019), modified by in-place landmark-guided vertical reflection of the eye and mouth regions (Alimohammadi, 2020) using dlib-based facial landmark detection (King, 2009; Kazemi & Sullivan, 2014) and presented upright and inverted (280 original–modified pairs; Appendix A). The Illumination task contains 96 scenes rendered in Kubric (Greff et al., 2022) each for cast-shadow, reflection, and light-direction interventions (288 pairs): for shadow and reflection conditions, the rendered floor region was mirrored to exchange shadows or reflections while objects remained fixed; for the light-direction condition, the central object was horizontally mirrored while the flanking objects and floor, including any cast shadows, remained unchanged. To produce graded item difficulty, we varied object geometry, orientation or color, height, material, and lighting parameters across scenes (Appendix A).

Human psychophysical experiments. Human observers rated the naturalness of each image on a 5-point scale (1: *very unnatural* to 5: *natural*) in online psychophysical experiments. Each trial presented a single image (Figure 1A). In the Thatcher task, each observer viewed upright and inverted faces from different identity sets, whereas illumination conditions were tested between subjects. In both tasks, original and modified versions of each scene were evaluated by the same observer across separate, randomized trials (Appendix B). Following response-consistency screening (143 observers retained for Thatcher, 105 for Illumination), each item received an average of 35.8 ratings in Thatcher and 26.2 in Illumination. Item-mean ratings $R(x)$ define two measures: single-image unnaturalness $a(x) = 5 - R(x)$ and the paired unnaturalness effect $h_i = R(x_i^{\text{orig}}) - R(x_i^{\text{mod}})$, which quantifies the

perceived unnaturalness introduced by each modification (i indexes stimulus pairs). Only item-level aggregates enter subsequent model analyses.

3.2 Generative Models and Native Loss Probing

We evaluate 25 open generative models (16 image, 9 video) across 11 families: (i) pixel-space image models (*JiT*, *PixelGen*, *HiDream-O1-Image*) (Li & He, 2026; Ma et al., 2026; Cai et al., 2026); (ii) latent-space image models (*Stable Diffusion v1.5/XL/3, FLUX.1/2, Qwen-Image*) (Rombach et al., 2022; Podell et al., 2024; Esser et al., 2024; Black Forest Labs, 2024; 2025; Wu et al., 2025); and (iii) spatiotemporal video models (*CogVideoX, LTX-Video, Wan2.1/2.2, HunyuanVideo*) (Yang et al., 2025; HaCohen et al., 2024; Team Wan et al., 2025; Kong et al., 2024). Static images are repeated as 17-frame clips, with 21 frames for CogVideoX1.5 to satisfy its temporal patching constraint. Text-conditioned models use the neutral or empty-prompt single-pass convention without classifier-free guidance; a control evaluation on a text-conditioned subset confirms that adding descriptive prompts does not alter our findings (Appendix O). Checkpoints, schedules, and evaluation specifications are listed in Appendix C.

Let $\ell_m(x, t, \epsilon)$ denote model m's native prediction loss (e.g., squared denoising error or flow-matching velocity loss) for input x, timestep t, and noise $\epsilon \sim \mathcal{N}(0, I)$. We aggregate over stored timesteps T_m (100 for image, 50 for video) and noise realizations E_m (20 draws):

$$L_m(x) = \frac{1}{|T_m||E_m|} \sum_{t \in T_m} \sum_{\epsilon \in E_m} \ell_m(x, t, \epsilon).$$

We use the average native prediction loss $L_m(x)$ as an empirical surrogate for negative log-likelihood. Because loss formulations, target spaces (pixel vs. latent), and latent dimensions differ across architectures, raw loss magnitudes $L_m(x)$ are not comparable across models. We evaluate two scores corresponding to human judgments: (i) the *unconditioned single-image loss* $L_m(x)$, compared against single-image unnaturalness $a(x)$ as a baseline control, and (ii) the *controlled paired loss change*

$$u_{m,i} = L_m(x_i^{\text{mod}}) - L_m(x_i^{\text{orig}}),$$

evaluated against the paired human unnaturalness effect h_i. A positive value ($u_{m,i} > 0$) indicates that the relational violation increased native denoising loss, consistent with human ratings where modifications reduce perceived naturalness ($h_i > 0$). Paired subtraction controls for shared scene content, focusing evaluation on the model's sensitivity to targeted relational violations.

3.3 Sensitivity and Alignment Metrics

To compare models across disparate raw loss scales, we evaluate two complementary dimensions. **Directional sensitivity** quantifies whether a model reliably penalizes relational violations on average, defined as the mean paired loss change divided by the model's sample standard deviation across all pairs within the same task: $S_{m,c} = \bar{u}_{m,c}/s_{m,\text{task}}$ where $\bar{u}_{m,c} = |I_c|^{-1} \sum_{i \in I_c} u_{m,i}$ averages over the pairs I_c in condition c; task-level sensitivity S_m pools the numerator across conditions. The denominator $s_{m,\text{task}}$ is computed across all pairs within the task ($N = 280$ for Thatcher, $N = 288$ for Illumination, using $N - 1$ degrees of freedom) and shared across conditions, which preserves relative response magnitudes. Human sensitivity is defined analogously from h_i. **Human alignment** measures whether graded loss increases track continuous human unnaturalness judgments across individual items, quantified by the item-level Pearson correlation $r_m = \text{corr}_i(u_{m,i}, h_i)$ within each task. Confidence intervals for human-dependent estimates use 10,000 participant-and-scene bootstrap resamples, whereas those for generative-model sensitivities use scene resampling alone. For cross-model sensitivity–alignment correlations, we additionally perform cluster bootstrap resampling over model families (Appendix E). To verify that associations between sensitivity and alignment are not driven by shared Monte Carlo sampling error, we estimated the two metrics from independent noise draws and observed closely matching results (Appendix P).

3.4 Baseline Models and Partial Correlation

We compare generative scores with a prespecified pool of 12 frozen image encoders (24M–8.1B parameters) and two five-metric IQA batteries (Appendix D). The encoder panel includes ResNet-50 (He et al., 2016), CLIP (Radford et al., 2021), DINOv2/v3 (Oquab et al., 2024; Sim'eoni et al.,

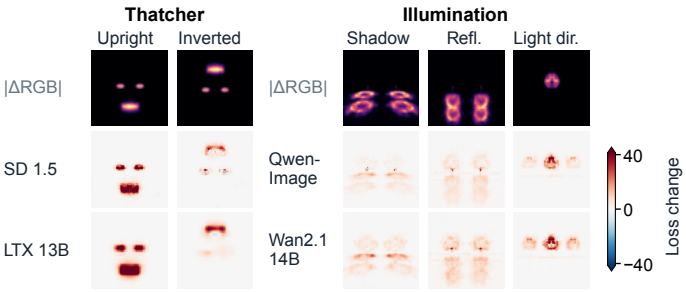

Figure 2: **Spatial loss changes.** Mean paired loss maps for two high-alignment models per task. Top-row pixel differences identify the modified regions. Complete 25-model maps: Appendix M.

2026), SigLIP 2 (Tschannen et al., 2025), PE-Core (Bolya et al., 2025), Qwen3-VL-Embedding (Li et al., 2026), and Jina CLIP v2 (Koukounas et al., 2024). Encoder scores use cosine distances between global vectors, or the mean distance between corresponding spatial patches for DINOv2/v3, capturing representational divergence across paired images. We evaluate representations across all intermediate blocks alongside each model's default final output, extracting features via fixed architecture-specific pooling (Appendix D). For each encoder and task, we select the candidate with the highest Pearson correlation with human judgments. While this post-hoc layer selection deliberately favors the encoder baselines over zero-shot generative scores, fixed feature distances measure the magnitude of change rather than the direction of degradation. Thus, they serve as a diagnostic baseline rather than an upper bound on encoder representations. The IQA batteries comprise FR-IQA (PSNR, SSIM, VIF, LPIPS, DISTS; Wang et al., 2004; Sheikh & Bovik, 2006; Zhang et al., 2018; Ding et al., 2022) and NR-IQA (NIQE, BRISQUE, HyperIQA, MUSIQ, LIQE; Mittal et al., 2013; 2012; Su et al., 2020; Ke et al., 2021; Zhang et al., 2023).

To test whether generative loss differences capture human-aligned naturalness judgments beyond standard image quality metrics and high-level visual encoders, we evaluate marginal alignment and bidirectional partial correlations against top-performing baselines. After selecting each encoder's best-performing layer, we assess marginal human alignment across candidate models, retaining the top five Generative models and top five Encoders alongside all five metrics from each prespecified IQA battery. For every selected generative–encoder pair (G, X), we compute $r(h, G \mid X)$ and $r(h, X \mid G)$, summarizing each 5×5 comparison by its medians. Top-five models are selected over each entire task and retained, together with the selected encoder layers, across its conditions (Appendix I). To account for selection uncertainty, confidence intervals are estimated via 10,000 participant-and-scene bootstrap resamples, repeating layer selection followed by model selection within each draw (Appendix E.3).

4 Results

4.1 Violation Sensitivity and Spatial Loss Changes

We first examined paired response profiles across representative stimulus manipulations (Figure 1B). In the Thatcher task, humans and top generative models strongly penalized upright facial feature inversions, with sensitivity dropping sharply under inversion (Figure 1B; 25-model median of 2.24 upright vs. .84 inverted; humans: 2.73 vs. 1.09; Appendix G). In the Illumination task, generative model losses tracked human unnaturalness across scene geometry: for cast shadows, both humans and models showed pronounced sensitivity to shadow rotations for objects with distinctive directional landmarks (e.g., teapot), but remained tolerant for shapes lacking prominent in-plane asymmetries (e.g., Spot; Figure 1B). Crucially, top vision encoders failed to capture this selective sensitivity, showing uniformly low responses across both objects (0.28 vs. 0.27), and similarly missed human sensitivity to light-direction changes on simple spheres (Figure 1B). Conversely, generative models were relatively insensitive to subtle reflection hue mismatches readily detected by humans, revealing a limit where native losses track continuous color differences rather than physical inconsistency.

Spatial loss-difference maps reveal where violation penalties arise across the scene (Figure 2). For upright Thatcher faces, loss increases concentrated around the inverted eyes and mouth, with marked attenuation under face inversion. For illumination interventions, loss increases were not confined to manipulated regions, extending from floor surfaces to objects directly above altered shadows or

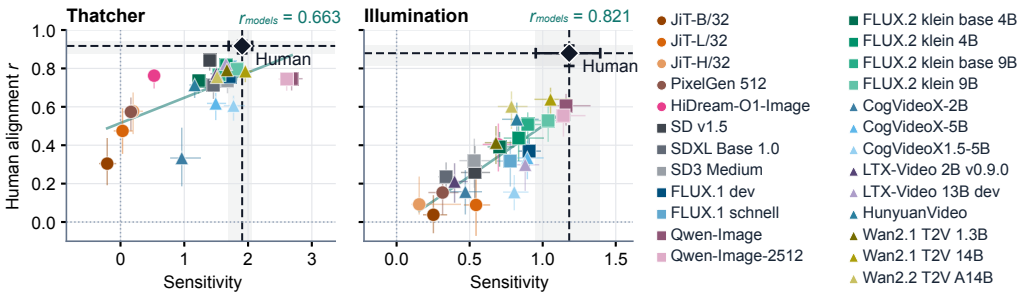

Figure 3: **Sensitivity and human alignment across models.** Directional sensitivity versus item-level human alignment for 25 generative models. Error bars and shaded bands indicate 95% bootstrap intervals (Section 3.3, Appendix E). Human diamonds and crosshairs mark sensitivity and mean split-half reliability.

reflections, and from the central object to unmodified flanking objects under light-direction shifts. These maps indicate that generative models evaluate contextual relationships across the scene rather than isolated local modifications.

4.2 Zero-Shot Alignment and Sensitivity Across Models

We next tested whether these qualitative patterns generalize to continuous item-level agreement across the full 25-model benchmark. Directional loss changes ($u_{m,i} = L_m(x_i^{\text{mod}}) - L_m(x_i^{\text{orig}})$) reliably tracked human unnaturalness judgments (h_i) across individual items (Figure 3, vertical axes), with median correlations of .747 on Thatcher and .334 on Illumination (peaking at .841 and .639). This alignment was not merely driven by coarse differences across conditions: after condition centering (subtracting condition-specific means from both model losses and human ratings), correlations remained positive for all 25 models on both tasks (medians of .411 and .271; peaking at .534 and .619; condition breakdowns in Appendix H). In contrast, single-image losses ($L_m(x)$) showed little correlation with human ratings (25-model medians of $-.20$ to .24; Appendix F.2), confirming that paired differencing is essential to separate relational naturalness from shared scene composition. Notably, alignment did not divide cleanly by generative formulation (diffusion versus flow), modeling space (pixel versus latent), or modality (image versus video), with substantial within-category variation. Furthermore, substantial human alignment was observed in models whose published training recipes do not report explicit preference optimization. On Thatcher, ImageNet-trained JiT exhibited alignment that scaled with model capacity (reaching $r = .577$ for JiT-H/32), alongside strong agreement in SD v1.5. On Illumination, Wan2.1 14B (Team Wan et al., 2025) achieved the highest overall correlation ($r = .639$, remaining .619 after condition centering), with solid alignment also sustained in video models like CogVideoX (Yang et al., 2025). These observations indicate that post-training preference optimization is not a prerequisite for the observed perceptual alignment.

Across the full tasks, directional sensitivity showed an overall positive association with item-level human alignment (Figure 3; Thatcher: $r = .663$, 95% CI $[.086, .905]$; Illumination: $r = .821$, $[.542, .922]$), showing that coarse violation responsiveness and fine-grained human alignment broadly co-vary across models. However, this coupling was not uniform across visual domains: positive associations were robust for cast shadows and inverted faces, but became weak or inconsistent for surface reflections and light direction (Appendices H and K).

Dissociation within denoising schedules. Despite this overall cross-model association, directional sensitivity and human alignment systematically dissociate within individual denoising schedules (Figure 4A,B). Sensitivity peaked later than human alignment in 20 of 25 models on Thatcher and all 25 on Illumination, with median within-model peak shifts (sensitivity minus alignment) of 0.141 and 0.444 on each model's normalized schedule ($[0, 1]$). Average violation responses and fine-grained human perceptual rankings thus peak at different noise levels during denoising (complete condition-specific profiles appear in Appendix N, Figures S29 and S30).

Relation to generation performance. To examine whether our loss-based diagnostic connects to practical generative capabilities, we evaluated human alignment against external generation

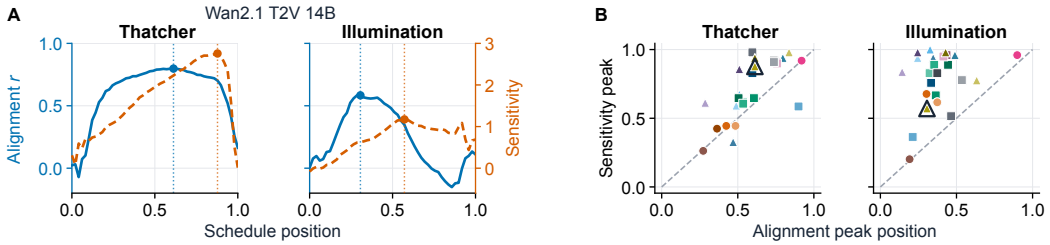

Figure 4: **Sensitivity and human alignment along denoising schedules.** (A) Representative schedule dynamics (Wan2.1 T2V 14B): human alignment r (blue, solid) and standardized sensitivity (orange, dashed) across normalized schedule positions ($0 \rightarrow 1$); solid dots mark peak positions. (B) Peak locations for all 25 models across normalized schedules ($[0, 1]$); the dashed diagonal marks equal timing. Model colors and marker shapes follow Figure 3; outlined markers identify Wan2.1 T2V 14B.

benchmarks, matching 11 image models to human-preference Elo (Artificial Analysis, 2026) and seven video models to VBench Quality (Huang et al., 2024). Even after controlling for directional sensitivity, partial Spearman correlations remained positive on Illumination for image models ($\rho = .670$) and both tasks for video models (.900 and .778; Appendix L). In contrast, the association on Thatcher images was weak ($\rho = .182$), which may reflect limited variation, as many models approach human-level agreement on faces (Figure 3). The reciprocal associations with sensitivity after controlling for alignment were weak across all four comparisons ($\rho = -.045$ to .218). While based on modest sample sizes, these exploratory associations suggest that loss-based perceptual alignment may reflect visual regularities that support practical generation quality.

4.3 Unique Perceptual Signals Beyond Vision and Quality Baselines

Across both tasks, top generative models achieved higher human alignment than the evaluated baselines (Figure 5A), even after selecting each encoder's best-performing layer (top-five median $r = .816$ vs. .660 on Thatcher; .602 vs. .492 on Illumination). In contrast, standard quality metrics showed substantially weaker correspondence: NR-IQA correlations remained near zero throughout (.087 on Thatcher, .009 on Illumination), arguing against generic degradation or manipulation artifacts, while weak FR-IQA results (.103 and .254) indicate that alignment is not driven merely by low-level change magnitude.

Generative scores also retained positive residual associations after controlling for encoder distances (Figure 5B). On Thatcher, generative scores provided incremental explanatory power beyond encoder distances (median partial correlation of .676, 95% CI $[.612, .701]$ vs. .130, $[.054, .228]$ for encoders), whereas on Illumination, both representations preserved non-redundant associations (.510 vs. .375). Broken down by condition, generative scores contributed distinct variance for upright faces, cast shadows, and light direction, while encoders retained larger associations for surface reflections (Figure S14, Appendix I). These results indicate that native generative loss landscapes and visual encoder representations capture complementary perceptual regularities.

5 Discussion

This study evaluated whether the native prediction losses of visual generative models reflect human naturalness perception under controlled relational modifications. Across diverse architectures, zero-shot loss differences track continuous human ratings and retain complementary perceptual structure beyond standard vision encoders and image quality metrics. While coarse violation sensitivity broadly covaries with perceptual alignment across models, sensitivity and fine-grained alignment systematically dissociate along denoising schedules and vary across visual domains. These findings suggest that current generative models internalize structured relational regularities that partially mirror the selective sensitivities and tolerances of human naturalness perception, rather than merely signaling physical violations or generic visual divergence. Beyond characterizing existing architectures, this

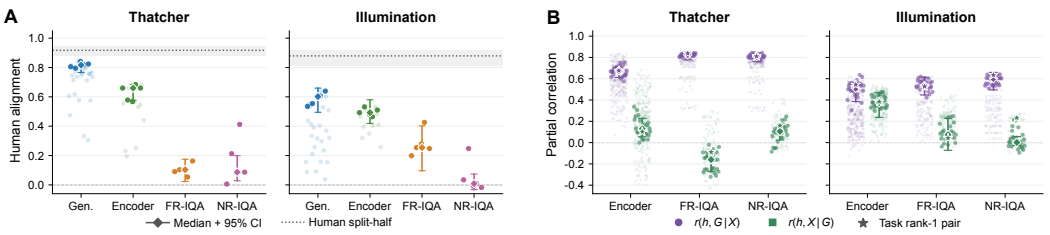

Figure 5: **Comparison with baselines.** (A) Model–human Pearson correlations. Saturated points mark top-five Generative and Encoder models and all IQA metrics; diamonds show medians. Dotted lines and bands show mean human split-half reliability and its 95% interval. (B) Bidirectional partial correlations ($r(h, G \mid X)$, purple circles; $r(h, X \mid G)$, green circles). Faint points show all pairs; saturated points show the selected 5×5 model pairs; stars mark the rank-1 pair. Large markers and bars in all panels show medians and 95% bootstrap CIs (Section 3.4, Appendix E.3).

diagnostic framework provides a foundation to track how generative representations converge toward or diverge from biological perception as models scale.

Orientation dependence in face perception. The orientation dependence of the Thatcher illusion is a classic hallmark of human face perception that deteriorates under inversion (Thompson, 1980; de Haas & Schwarzkopf, 2018). This orientation dependence arises in feedforward representations (Jacob et al., 2021) and in our encoder baselines (Figure 1B). Paralleling this, generative diffusion and flow models exhibit pronounced orientation-dependent loss penalties across diverse architectures. Notably, the ImageNet-trained JiT models (Li & He, 2026) demonstrate that this orientation penalty can arise without training on a dedicated face dataset, scaling systematically with model capacity (Appendix G). Generative losses also tracked naturalness variations within upright faces alone (median $r = .493$), paralleling findings that generative evidence tracks human facial attractiveness (López & Triesch, 2026). Whereas accounting for face illusions has often relied on structured 3D inverse-graphics models (Yildirim et al., 2020), our findings indicate that 2D generative loss landscapes naturally mirror orientation-dependent perceptual sensitivity.

Physical scene illumination. Human sensitivity to shadow and reflection inconsistencies varies with spatial configuration (Farid & Bravo, 2010; Nightingale et al., 2019), consistent with correspondence accounts where salient directional landmarks (e.g., teapot spout and handle) enable feature matching that ambiguous shapes do not (Casati, 2008; Norman et al., 2009). Beyond the representative comparisons in Figure 1B, factor-level profiles across shadow and reflection rotations as well as light direction across object shapes confirm that generative losses systematically track human sensitivity to these physical scene manipulations (Appendix J, Figures S16–S18). The weaker model response to small reflection hue shifts (Figure 1B) indicates that native losses track continuous low-level color differences rather than the qualitative break in expected physical relationships that dominates human perception. Loss formulations defined over more human-aligned perceptual representations might help narrow this divergence.

Dynamics along denoising schedules. Directional sensitivity typically peaked later than human alignment, showing that the strongest loss penalties do not always yield the best agreement with fine-grained human perception. As denoising schedules progress from global composition to fine details (Sclocchi et al., 2025; Tınaz et al., 2025; G"org"un et al., 2026), human alignment on relational tasks may peak mid-schedule when coarse scene structure is resolved, whereas later stages emphasize local details that inflate loss penalties without improving perceptual agreement.

Limitations. First, our psychophysical intervention framework isolates relational regularities at the expense of scene diversity; extending controlled interventions to unconstrained scenes remains challenging without introducing low-level artifacts. Second, off-the-shelf comparisons cannot fully isolate core generative modeling from model scale or post-training recipes, nor can they disentangle generic visual statistics from the human-curated, preference-favored distributions inherent to modern training corpora. Third, our observational analyses cannot establish whether directly optimizing alignment causally improves generation quality. Finally, sensitivity–alignment coupling varies across visual domains (e.g., surface reflections vs. cast shadows); clarifying how training data and objectives shape these differences remains an open question.

AI USE STATEMENT

Generative AI tools were used at several stages of the research and writing process. They were used to identify potentially relevant prior work and related studies; to discuss and obtain feedback on aspects of research methodology, analysis strategy, and interpretation of results; to assist with data analysis and plotting code; and to draft, revise, and polish portions of the manuscript, including LaTeX formatting. Candidate references and related work identified with AI assistance were checked against the original sources, and AI-assisted code and analyses were inspected and verified by the authors.

In addition, an image generation model was used to create the synthetic face image shown in Figures 1 and S1 solely for illustrative purposes to depict the stimulus manipulation. This synthetic image was not used in the human experiments or model analyses.

All research decisions, analyses, interpretations, and claims were made and verified by the authors. All AI-assisted text, code, and other outputs were reviewed and edited as appropriate, and the authors take full responsibility for the contents of this work.

ETHICS STATEMENT

The human behavioral experiments were approved by the Institutional Review Board (Ethics Committee) of the authors' institution. All participants were adults and provided informed consent via an electronic consent form prior to participation. Detailed participant demographics, experimental procedures, and task instructions are provided in Appendix B.

REPRODUCIBILITY STATEMENT

The supplementary material provides model-scoring code, configurations, stimulus images, model scores, anonymized participant ratings, and code for screening, bootstrap inference, and the analyses listed in its README. These materials will be made publicly available upon publication. Model specifications and experimental and statistical procedures are described in Section 3 and Appendices A–E.

REFERENCES

Erfan Alimohammadi. Thatcher Effect Dataset Generator. https://github.com/Erfaniaa/thatcher-effect-dataset-generator, 2020. Open-source software under GNU General Public License v3.0.

Artificial Analysis. Image generation benchmarking methodology. https://artificialanalysis.ai/image/methodology, 2026. Accessed September 12, 2026.

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. VideoPhy: Evaluating physical commonsense for video generation. In *International Conference on Learning Representations*, 2025.

Black Forest Labs. FLUX. https://github.com/black-forest-labs/flux, 2024.

Black Forest Labs. FLUX.2: Frontier visual intelligence. https://bfl.ai/blog/flux-2, 2025.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Bangalath, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Shang-Wen Li, Piotr Doll'ar, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Qi Cai et al. HiDream-O1-Image: A natively unified image generative foundation model with pixel-level unified transformer. *arXiv preprint arXiv:2605.11061*, 2026.

Roberto Casati. The copycat solution to the shadow correspondence problem. *Perception*, 37(4): 495–503, 2008. doi: 10.1068/p5588.

Patrick Cavanagh. The artist as neuroscientist. *Nature*, 434(7031):301–307, 2005.

Chaofeng Chen and Jiadi Mo. IQA-PyTorch: PyTorch toolbox for image quality assessment. https://github.com/chaofengc/IQA-PyTorch, 2022.

Kevin Clark and Priyank Jaini. Text-to-image diffusion models are zero-shot classifiers. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46433–46449, 2023.

Benjamin de Haas and D. Samuel Schwarzkopf. Feature-location effects in the Thatcher illusion. *Journal of Vision*, 18(4):16, 2018.

Joshua R. de Leeuw, Rebecca A. Gilbert, and Bj"orn Luchterhandt. jsPsych: Enabling an open-source collaborative ecosystem of behavioral experiments. *Journal of Open Source Software*, 8(85):5351, 2023.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2022. doi: 10.1109/TPAMI.2020.3045810.

Zhengfang Duanmu, Wentao Liu, Zhongling Wang, and Zhou Wang. Quantifying visual image quality: A Bayesian view. *Annual Review of Vision Science*, 7:437–464, 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas M"uller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633, 2024.

Hany Farid. Lighting (in)consistency of paint by text. *arXiv preprint arXiv:2207.13744*, 2022. URL https://arxiv.org/abs/2207.13744.

Hany Farid and Mary J Bravo. Image forensic analyses that elude the human visual system. In Nasir D Memon, Jana Dittmann, Adnan M Alattar, and Edward J Delp, III (eds.), *Media Forensics and Security II*, volume 7541, pp. 52–61. SPIE, 4 February 2010. doi: 10.1117/12.837788.

Honghao Fu, Yufei Wang, Wenhan Yang, Alex C. Kot, and Bihan Wen. DP-IQA: Utilizing diffusion prior for blind image quality assessment in the wild. *arXiv preprint arXiv:2405.19996*, 2024.

Alexandra G'omez-Villa, Adri'an Mart'ın, Javier Vazquez-Corral, Marcelo Bertalm'ıo, and Jes'us Malo. Color illusions also deceive CNNs for low-level vision tasks: Analysis and implications. *Vision Research*, 176:156–174, 2020.

Alexandra Gomez-Villa, Kai Wang, C. Alejandro Párraga, Bartłomiej Twardowski, Jesús Malo, Javier Vazquez-Corral, and Joost van de Weijer. The art of deception: Color visual illusions and diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18642–18652, 2025.

Ada G"org"un, Fawaz Sammani, Nikos Deligiannis, Bernt Schiele, and Jonas Fischer. Temporal concept dynamics in diffusion models via prompt-conditioned interventions. In *International Conference on Learning Representations*, 2026.

Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J. Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam Laradji, Hsueh-Ti (Derek) Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, Cengiz Oztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi, Fangcheng Zhong, and Andrea Tagliasacchi. Kubric: A scalable dataset generator. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3749–3761, 2022.

Yoav HaCohen et al. LTX-Video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS 2020*, 2020.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Huang_VBench_Comprehensive_Benchmark_Suite_for_Video_Generative_Models_CVPR_2024_paper.html.

Georgin Jacob, R T Pramod, Harish Katti, and S P Arun. Qualitative similarities and differences in visual object representations between brains and deep networks. *Nature Communications*, 12(1): 1872, 25 March 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-22078-3.

Hamidreza Kamkari, Brendan Leigh Ross, Jesse C. Cresswell, Anthony L. Caterini, Rahul Krishnan, and Gabriel Loaiza-Ganem. A geometric explanation of the likelihood OOD detection paradox. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 22908–22935, 2024.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1867–1874, 2014.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *IEEE/CVF International Conference on Computer Vision*, 2021.

Davis E King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10: 1755–1758, 2009.

Diederik P. Kingma and Ruiqi Gao. Understanding diffusion objectives as the ELBO with simple data augmentation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 70932–70956, 2023.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Why normalizing flows fail to detect out-of-distribution data. In *Advances in Neural Information Processing Systems*, volume 33, pp. 20578–20589, 2020.

Weijie Kong et al. HunyuanVideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Andreas Koukounas et al. jina-clip-v2: Multilingual multimodal embeddings for text and images. *arXiv preprint arXiv:2412.08802*, 2024.

Eric Cooper Larson and Damon Michael Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19(1):011006, January 2010. ISSN 1017-9909,1560-229X. doi: 10.1117/1.3267105.

Mingxin Li et al. Qwen3-VL-Embedding and Qwen3-VL-Reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36115–36125, 2026.

Francisco M. López and Jochen Triesch. Beauty is in the ELBO of the beholder: A variational account of processing fluency in face perception. *arXiv preprint arXiv:2608.24219*, 2026.

Zehong Ma, Ruihan Xu, and Shiliang Zhang. PixelGen: Improving pixel diffusion with perceptual supervision. *arXiv preprint arXiv:2602.02493*, 2026.

Pascal Mamassian. Impossible shadows and the shadow correspondence problem. *Perception*, 33 (11):1279–1290, 1 November 2004. ISSN 0301-0066,1468-4233. doi: 10.1068/p5280.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012.

Anish Mittal, Rajiv Soundararajan, and Alan Conrad Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 948–958, 2026.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *International Conference on Learning Representations*, 2019.

Sophie J Nightingale, Kimberley A Wade, Hany Farid, and Derrick G Watson. Can people detect errors in shadows and reflections? *Attention, Perception, & Psychophysics*, 81(8):2917–2943, 28 November 2019. ISSN 1943-3921,1943-393X. doi: 10.3758/s13414-019-01773-w.

J Farley Norman, Young-lim Lee, Flip Phillips, Hideko F Norman, L RaShae Jennings, and T Ryan McBride. The perception of 3-D shape from shadows cast onto curved surfaces. *Acta Psychologica*, 131(1):1–11, 2009. doi: 10.1016/j.actpsy.2009.01.007.

Maxime Oquab, Timoth'ee Darcet, Th'eo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Yuri Ostrovsky, Patrick Cavanagh, and Pawan Sinha. Perceiving illumination inconsistencies in scenes. *Perception*, 34(11):1301–1314, 2005. ISSN 0301-0066. doi: 10.1068/p5418.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas M"uller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763. PMLR, 2021.

Simon Raviv and Gal Chechik. Assessing image quality using a simple generative representation. *arXiv preprint arXiv:2404.18178*, 2024.

Jie Ren, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark A. DePristo, Joshua V. Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/1e79596878b2320cac26dd792a6c51c9-Abstract.html.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Michele A Saad, Alan Conrad Bovik, and Christophe Charrier. Blind image quality assessment: A natural scene statistics approach in the DCT domain. *IEEE Transactions on Image Processing*, 21 (8):3339–3352, 2012.

Srijani Saha, Talia Konkle, and George A Alvarez. A unified account of lightness illusions via edge-based reconstruction of natural images. *bioRxiv preprint*, 2026. doi: 10.64898/2026.04.07.716245. URL https://www.biorxiv.org/content/10.64898/2026.04.07.716245v2.

Ayush Sarkar, Hanlin Mai, Amitabh Mahapatra, Svetlana Lazebnik, David A. Forsyth, and Anand Bhattad. Shadows don't lie and lines can't bend! generative models don't know projective geometry... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28140–28149, 2024.

Antonio Sclocchi, Alessandro Favero, and Matthieu Wyart. A phase transition in diffusion models reveals the hierarchical nature of data. *Proceedings of the National Academy of Sciences*, 122(1): e2408799121, 2025.

Joan Serrà, David Àlvarez, Vicenç Gómez, Olga Slizovskaia, José F. Núñez, and Jordi Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SyxIWpVYvr.

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. In *Conference on Language Modeling*, 2024.

Hamid R Sheikh and Alan C Bovik. Image information and visual quality. *IEEE Transactions on Image Processing*, 15(2):430–444, 2006. doi: 10.1109/TIP.2005.859378.

B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London, 1986.

Oriane Sim'eoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Kwang Moo Yi, Michael Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timoth'ee Darcet, Th'eo Moutakanni, Florian Sentana, Mike Roberts, Andrea Vedaldi, Metin Tolan, John Brandt, Camille Couprie, Julien Mairal, Herv'e J'egou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *Transactions on Machine Learning Research*, 2026.

Yiren Song, Xiaokang Liu, and Mike Zheng Shou. DiffSim: Taming diffusion models for evaluating visual similarity. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16904–16915, 2025.

Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Team Wan et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Peter Gage Thompson. Margaret Thatcher: A new illusion. *Perception*, 9(4):483–484, 1980. doi: 10.1068/p090483.

Berk Tınaz, Zalan Fabian, and Mahdi Soltanolkotabi. Emergence and evolution of interpretable concepts in diffusion models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 166943–166986, 2025.

Michael Tschannen et al. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

John D Wilder, Wendy J Adams, and Richard F Murray. Shape from shading under inconsistent illumination. *Journal of Vision*, 19(6):2, 3 June 2019. ISSN 1534-7362. doi: 10.1167/19.6.2.

Chenfei Wu et al. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

You-Zhe Xie, Yu-Hsuan Li, Jie-Ying Lee, Kaipeng Zhang, Yu-Lun Liu, and Zhixiang Wang. YoCausal: How far is video generation from world model? a causality perspective. *arXiv preprint arXiv:2605.30346*, 2026. URL https://arxiv.org/abs/2605.30346.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, 2025.

Ilker Yildirim, Mario Belledonne, Winrich Freiwald, and Josh Tenenbaum. Efficient inverse graphics in biological face processing. *Science Advances*, 6(10):eaax5979, 2020.

Jianhao Yuan, Fabio Pizzati, Francesco Pinto, Lars Kunze, Ivan Laptev, Paul Newman, Philip Torr, and Daniele De Martini. LikePhys: Evaluating intuitive physics understanding in video diffusion models via likelihood preference. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=6UJf6B8RZ8.

Alan Yuille and Daniel Kersten. Vision as Bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences*, 10(7):301–308, 2006.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

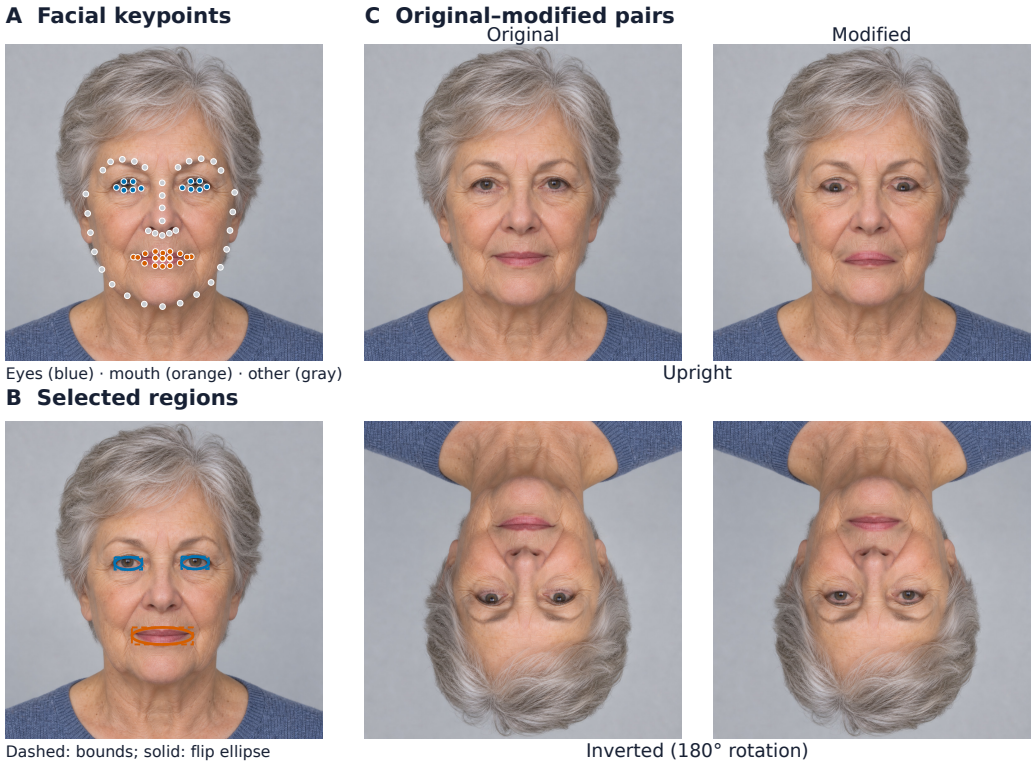

Figure S1: **Thatcher stimulus construction.** **(A)** The 68 facial landmarks detected on an AI-generated fictional face (eyes: blue; mouth: orange; remaining points: gray). **(B)** Selected regions for feature inversion. Dashed rectangles indicate padded landmark bounds; solid ellipses delimit vertically inverted regions feathered along their boundaries. **(C)** Original (left) and Thatcherized (right) versions in upright (top) and inverted (bottom; exact 180° rotation) orientations. The face is used only to illustrate the procedure and was not included in the human experiments or model analyses.

A Stimulus Construction

A.1 Thatcher Faces.

The stimulus set comprises 140 aligned 1024×1024 face identities, each evaluated across four conditions: original upright, Thatcherized upright, original inverted, and Thatcherized inverted (560 images total; construction procedure illustrated in Figure S1).

Source dataset and license filtering. Source face images were sampled from the Flickr-Faces-HQ (FFHQ) dataset (Karras et al., 2019), which provides aligned 1024×1024 portraits. Source images were restricted to entries marked CC0 1.0 or Public Domain Mark 1.0 in the official FFHQ metadata (`ffhq-dataset-v2.json`). From this pool, we chose 140 distinct public-domain face identities for the experimental stimuli, and an additional 5 disjoint identities for practice trials (10 practice trials presenting original and modified pairs across orientations).

Feature inversion and boundary blending. To generate Thatcherized faces, we adapted the open-source pipeline by Alimohammadi (2020) (released under the GNU General Public License v3.0). For each face image, 68 facial landmarks were detected using `dlib` (King, 2009) with its pre-trained shape predictor (Kazemi & Sullivan, 2014) (Figure S1A). Landmarks corresponding to the left eye (points 36–41), right eye (points 42–47), and mouth (points 48–67) were used to define bounding regions with modest padding offsets around each feature (Figure S1B, dashed rectangles). Within each feature region, pixels inside an elliptical mask were vertically inverted in place across the

feature's horizontal midline (Figure S1B, solid ellipses). To minimize high-contrast seam artifacts and artificial edges, Gaussian blending feathered the transition along the elliptical perimeter into the surrounding skin. Inverted stimuli were generated by rotating both the original and Thatcherized images by 180° (Figure S1C).

A.2 Illumination scenes.

We generated synthetic 1024×1024 stimuli using Kubric (Greff et al., 2022) across three physical operations: cast shadows (*Shadow*), surface reflections (*Reflection*), and lighting direction (*Light direction*). Each operation comprised 48 base configurations. Horizontally mirroring each base pair produced another 48 pairs, yielding 96 pairs per operation and 288 pairs in total.

Horizontal-flip manipulations. In *Shadow* and *Reflection*, we horizontally flipped the floor region while leaving the objects unchanged, exchanging their shadows or reflections (Figure S2). The two objects were placed symmetrically about the image's vertical midline, so the flip mapped their positions onto one another without introducing a positional offset. Additionally, objects were elevated above the floor (see below) to ensure a spatial gap; if objects contacted the ground, swapping the floor would alter contact contours or introduce spurious edge discontinuities at the object base, confounding loss differences with low-level edge artifacts. In *Light direction*, we horizontally flipped the central object while leaving the floor and flanking objects unchanged. This reversed its shading relative to the flanking objects and, in low-height scenes, disrupted the correspondence between its shading and its unchanged cast shadow. Every intervention rearranged existing pixels without adding newly rendered visual content.

Scene factors. Figures S3–S8 illustrate the stimuli across all 44 analysis levels evaluated in Appendix J. To present these levels concisely without displaying every factorial combination, object variations (shape, rotation, and color) are shown under shared baseline settings, while separate pairwise panels show both levels of each environmental factor (such as height, shadow softness, or surface roughness).

For *Shadow*, six Shape pair cases (S1) and six Rotation cases (S2) were each crossed with *Height* (S3: low, 0.1; high, 0.3) and *Shadow softness* (S4: soft or hard), giving $(6 + 6) \times 2 \times 2 = 48$ base configurations. Soft shadows have blurred edges and were produced by a large area light; hard shadows have sharp edges and were produced by a small, point-like light source.

For *Reflection*, four Shape pair cases (R1), four Rotation cases (R2), and four Color cases (R3) were each crossed with Height (R4: low, 0.1; high, 0.3) and *Floor roughness* (R5: smooth, 0.0; rough, 0.3), giving $(4 + 4 + 4) \times 2 \times 2 = 48$ configurations. The four Color cases combine two shapes, cube and torus, with a small or large hue shift (0.05 or 0.5 turns on the hue circle). Smooth floors produce mirror-like reflections; rough floors blur them. Shape pair, Rotation, and Color are separate sets of cases, not mutually crossed factors.

For *Light direction*, six levels of *Shape* (L1: cube, cylinder, sphere, Spot, teapot, and torus knot) were fully crossed with *Material* (L2: diffuse or metal), Height (L3: low, 0.2; high, 2.0), and *Light angle* (L4: 15° or 30°), giving $6 \times 2 \times 2 \times 2 = 48$ configurations. Cast shadows were disabled in the high condition. Heights are expressed in renderer scene units.

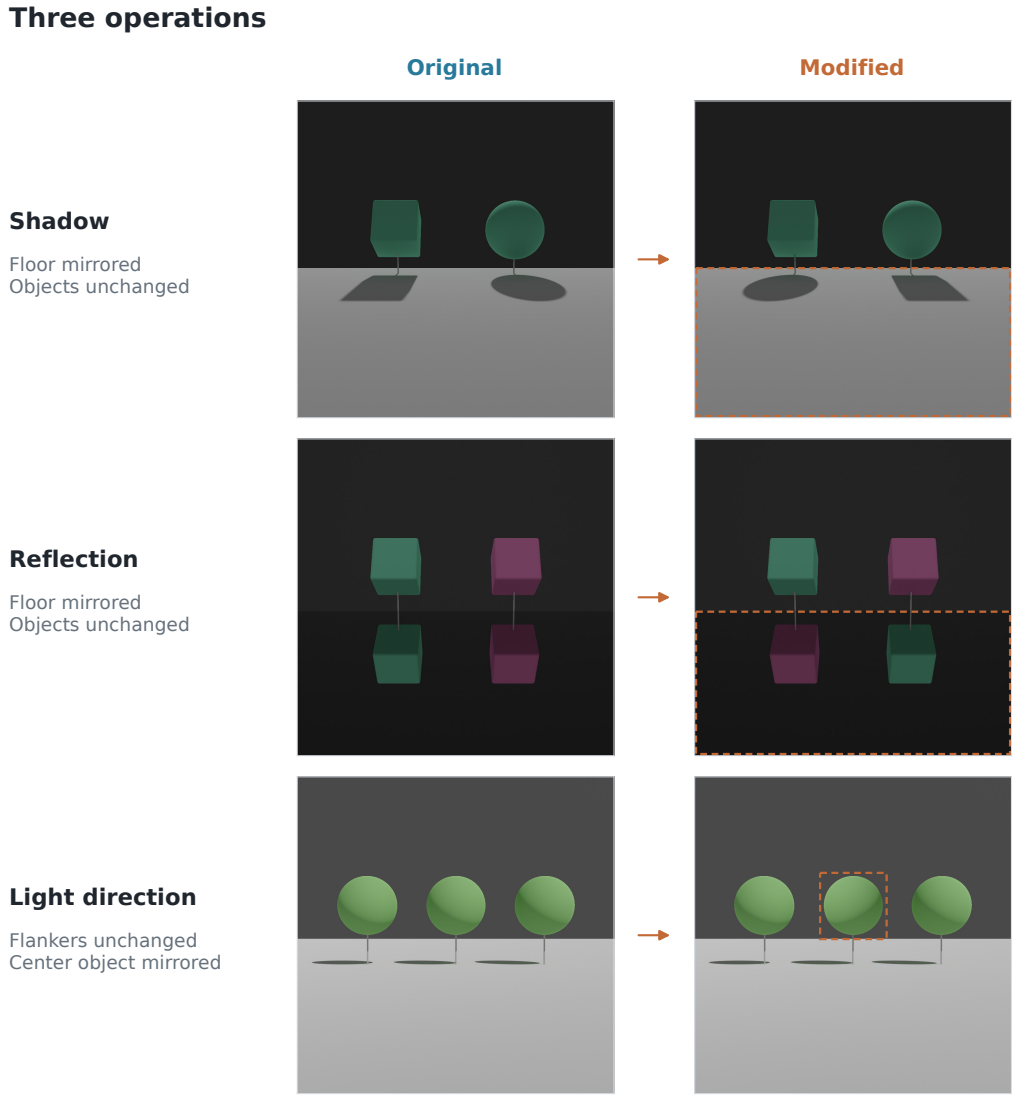

Figure S2: **Illumination stimulus construction.** Original (left) and modified (right) images for the three operations. The floor region is horizontally flipped in Shadow and Reflection, and the central object in Light direction. Orange outlines indicate the affected regions.

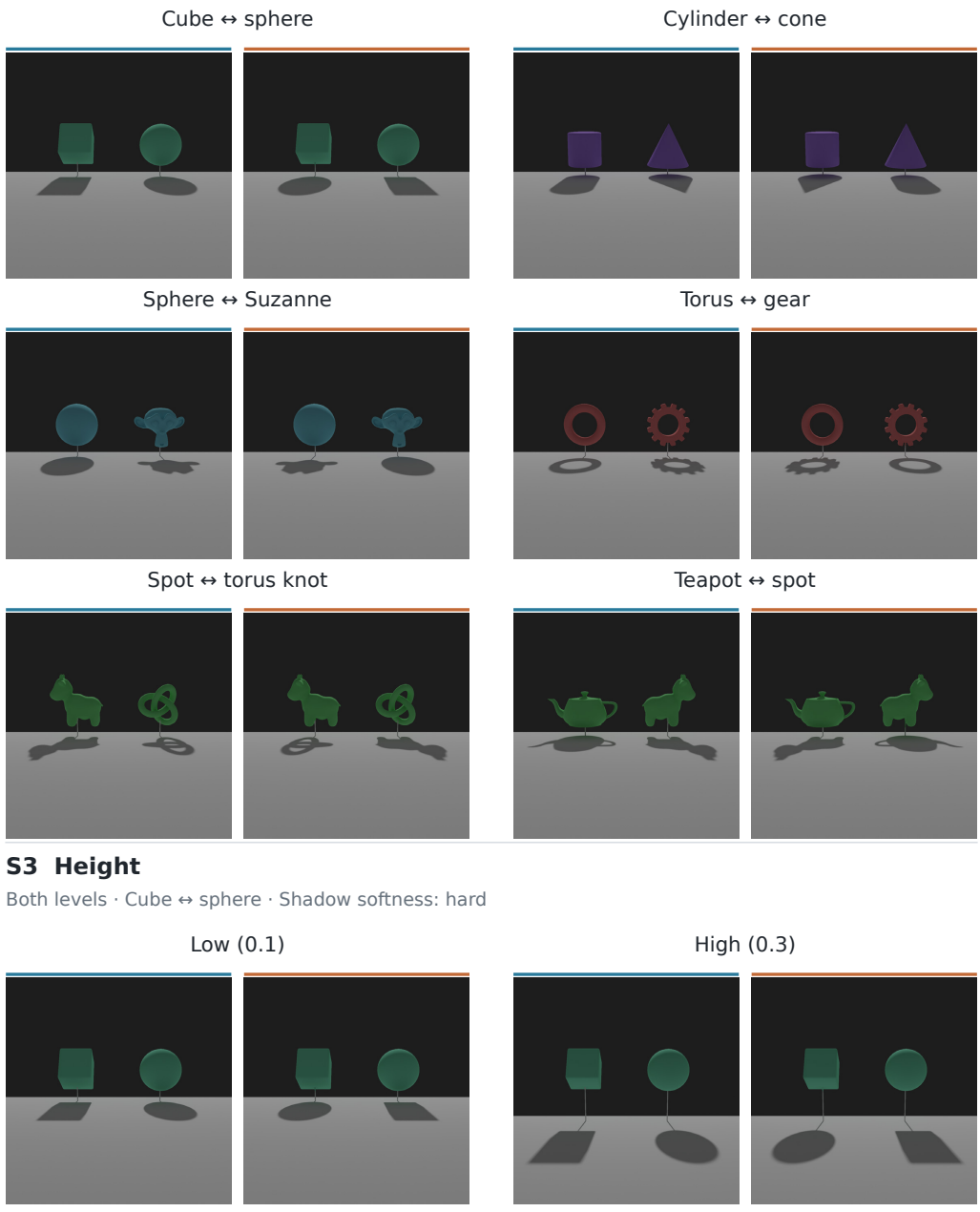

Figure S3: **Shadow: shape pairs and height.** Each example pairs the full original image (left, blue rule) with its modification (right, orange rule). S1 exchanges shadows between distinct object shapes to introduce salient, categorical geometric violations. S3 compares both object heights for the cube–sphere pair, testing whether elevating objects above the floor reduces the salience of shadow inconsistencies.



Figure S4: **Shadow: rotations and shadow softness.** S2 shows all six orientation pairs, with angles indicating object orientations. In modified images, exchanging shadows alters directional correspondence: this inconsistency is subtle for shapes lacking distinct landmarks (e.g., torus knot) or pairs with near-symmetric 2D silhouettes despite 3D depth asymmetry (e.g., Spot $-45° \leftrightarrow 135°$), requiring observers to scrutinize depth relationships, whereas prominent directional features (e.g., teapot spout) make it immediately obvious. S4 compares soft and hard shadows, testing whether blurred boundaries increase tolerance to discrepancies.

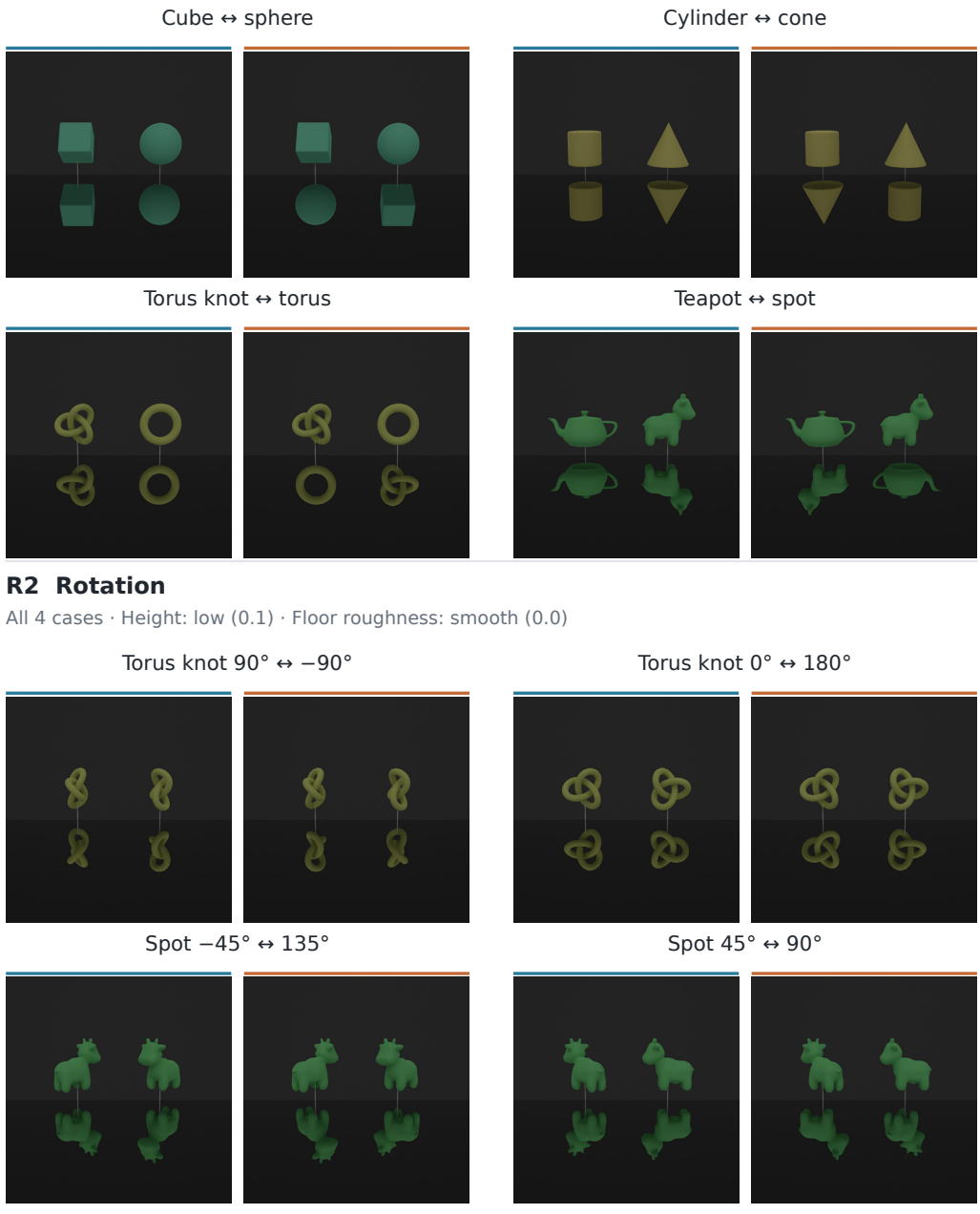

Figure S5: **Reflection: shape pairs and rotations.** R1 shows all four shape pairs and R2 all four orientation pairs, with height and floor roughness held at the baseline settings. R1 introduces categorical shape discrepancies between objects and their reflections, whereas R2 tests sensitivity to mirror-reflection symmetry across 3D orientations. In R2, inconsistencies are subtle and difficult to detect for shapes lacking distinctive landmarks (e.g., torus knot) or pairs appearing near-symmetric in 2D (e.g., Spot $-45^\circ \leftrightarrow 135^\circ$), whereas salient directional asymmetry (e.g., Spot $45^\circ \leftrightarrow 90^\circ$) makes violations apparent.

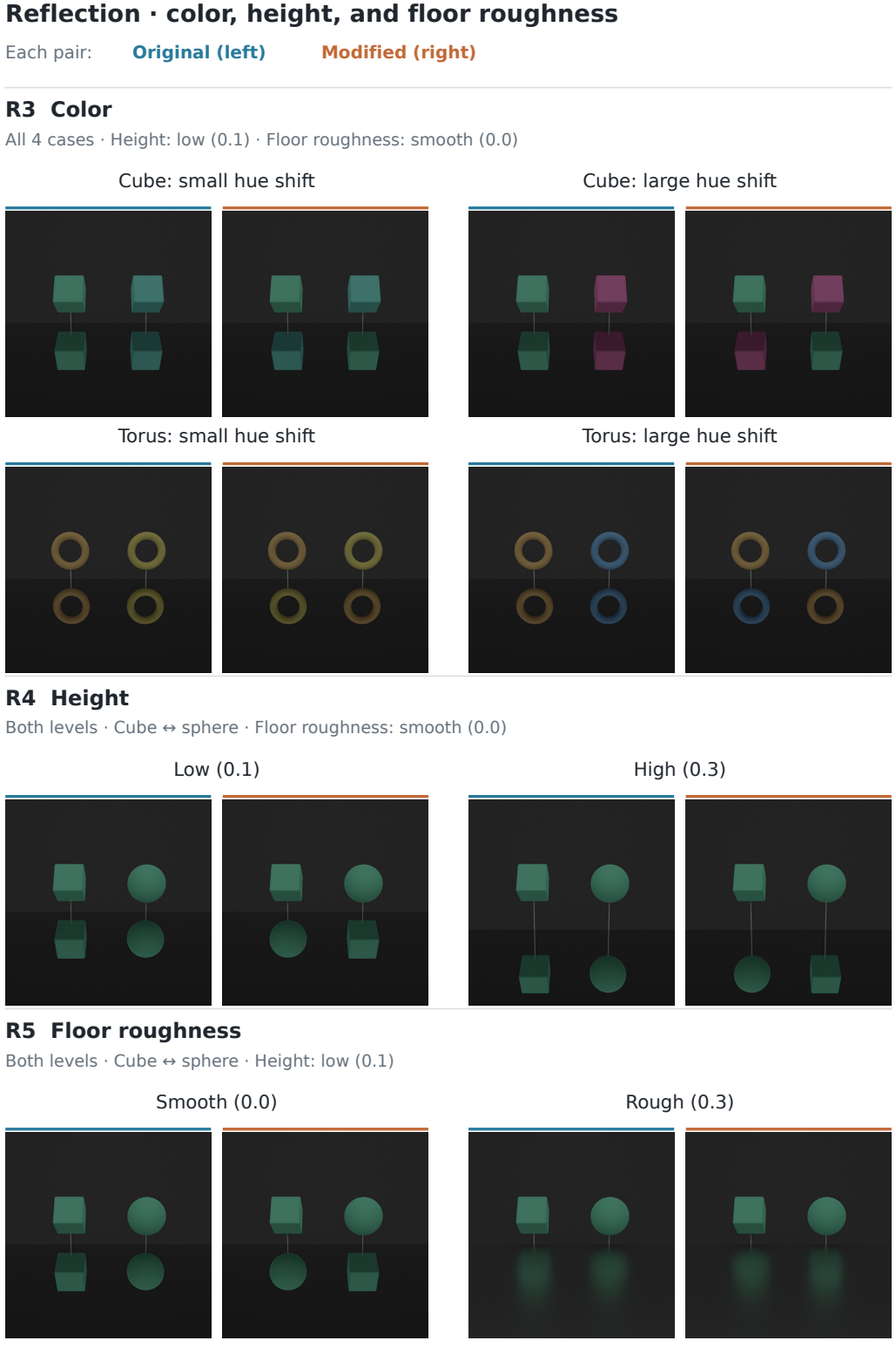

Figure S6: **Reflection: color, height, and floor roughness.** R3 introduces chromatic mismatches without altering geometry, testing sensitivity to color differences. R4 and R5 compare heights and surface roughness, examining how spatial proximity and blur modulate the salience of reflection anomalies.

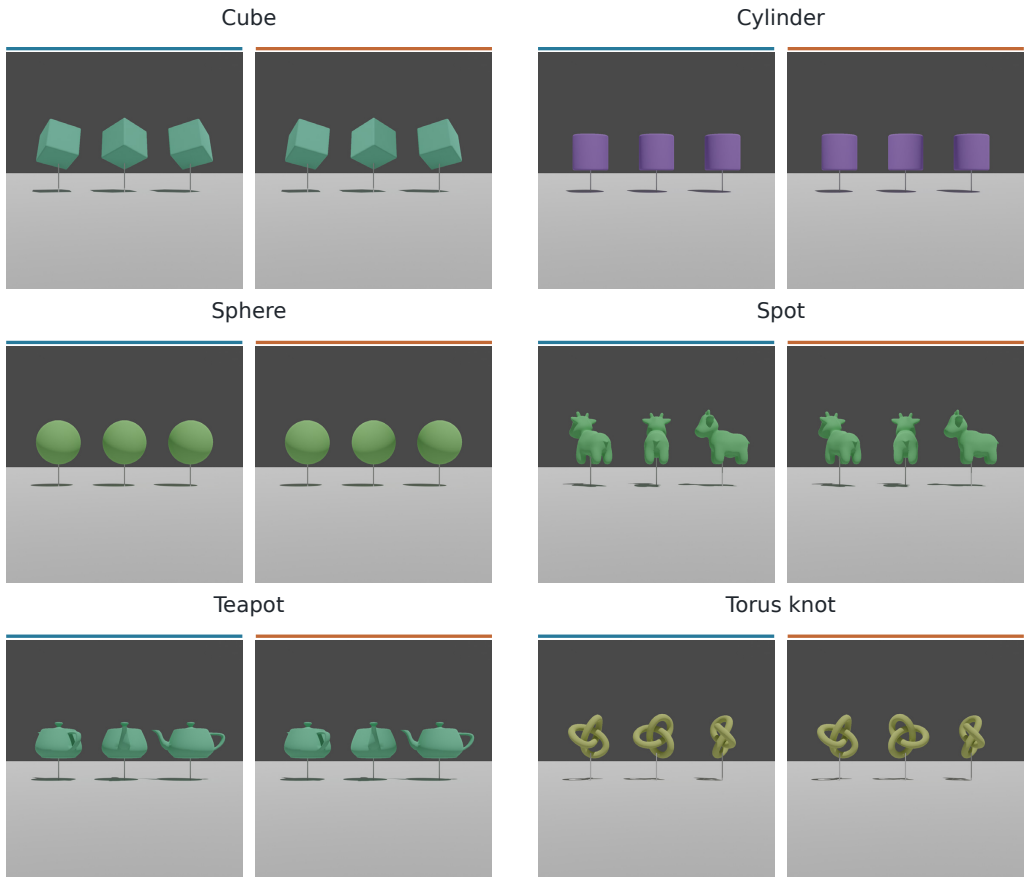

Figure S7: **Light direction: shapes.** L1 shows every shape at shared baseline settings. Horizontally mirroring only the central object reverses its shading relative to the flanking objects and unchanged cast shadows. This tests how object geometry affects lighting consistency, contrasting simple shapes (e.g., sphere) where illumination direction is unambiguous with complex shapes where lighting discrepancies are easily overlooked.

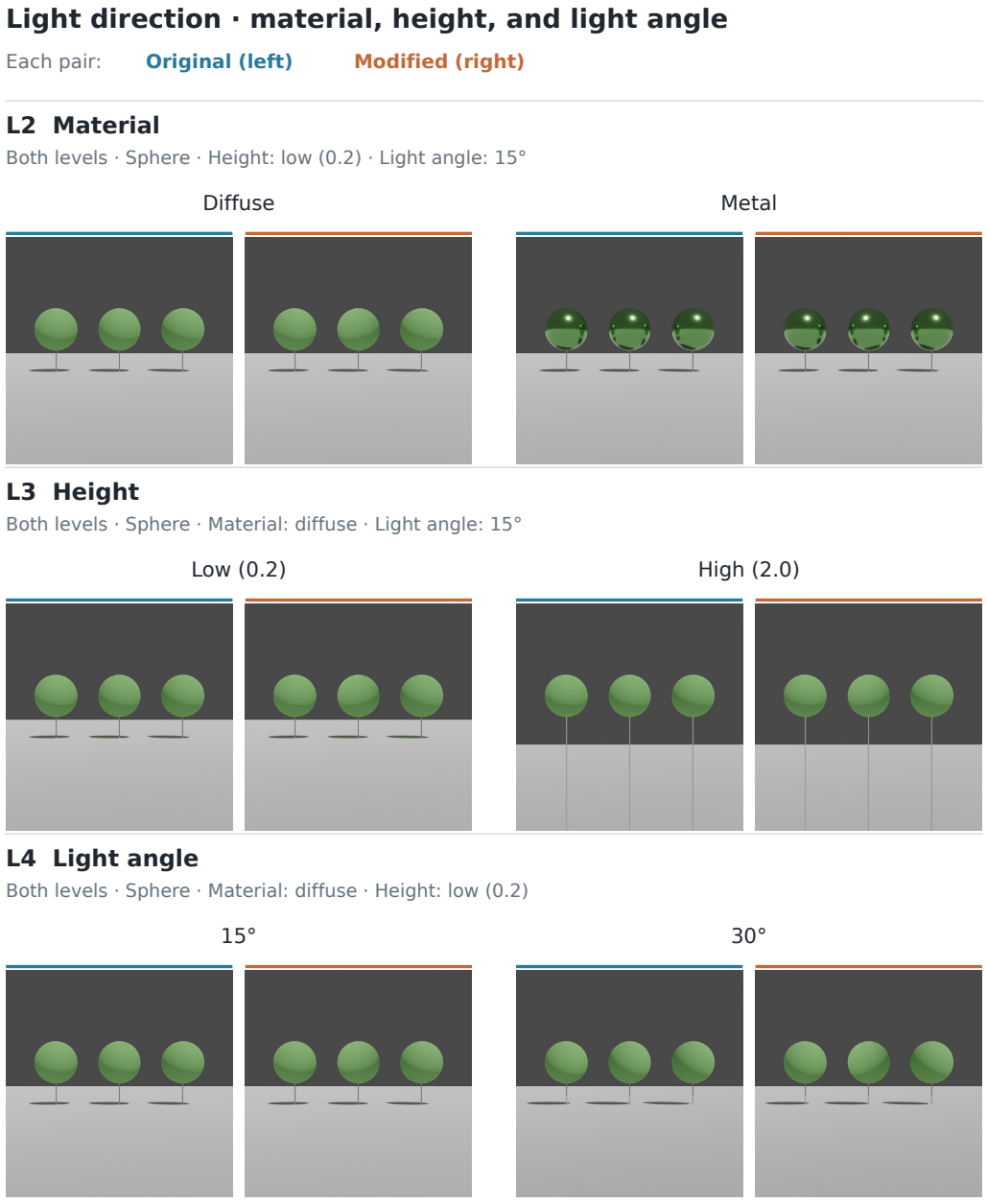

Figure S8: **Light direction: material, height, and light angle.** Each factor is illustrated at both levels using spheres, holding other factors fixed. L2 tests how surface material properties modulates lighting perception (diffuse shading gradients vs. specular highlights). L3 examines whether the presence of a consistent cast shadow strengthens the perceived lighting conflict compared to floating objects without shadows. L4 tests the effect of illumination inclination (15° vs. 30°).

B HUMAN EXPERIMENT DETAILS

B.1 APPARATUS AND DISPLAY INTERFACE

Online psychophysical experiments were developed using jsPsych (v7) (de Leeuw et al., 2023) and hosted on a dedicated web server. Prior to starting the experiment, participants were required to switch their browsers to full-screen mode, and automated checks verified compatible rendering environments. Stimuli were displayed in the center of the screen at 512×512 pixels against a neutral background.

Participants provided naturalness ratings on a discrete 5-point horizontal slider with text labels: 1 (*Very unnatural*), 2 (*Unnatural*), 3 (*Slightly unnatural*), 4 (*Moderately natural*), and 5 (*Natural*). To avoid anchoring effects, the slider handle was initialized to a random integer position (1–5) on every trial. Participants had to click or move the handle before the continue button was enabled, so that retaining the initial position required an active response. Stimuli remained on screen without time limits until submitted, followed by a 500 ms blank inter-trial interval (ITI).

B.2 TASK INSTRUCTIONS

Because the stimulus manipulations introduced subtle violations of configural or physical regularities within otherwise realistic scenes, relying solely on an unguided, generic query of "naturalness" risks inattentive oversights or idiosyncratic evaluation criteria (e.g., judging superficial photographic aesthetics rather than physical plausibility). To ensure that participants attended to the critical structural attributes and to calibrate rating criteria across observers, participants were familiarized with the candidate manipulation types through illustrated instructions and visual examples prior to the experiment:

- **Thatcher task:** Participants were instructed: *"Rate the naturalness of the facial appearance on a 5-point scale. Internal facial parts (eyes and mouth) may have artificial image manipulations applied. If you perceive a natural face without manipulation, select 5 (Natural). If you feel unnaturalness due to manipulation, select 1–4 according to the degree of unnaturalness. Faces may also appear upside down; please evaluate them using the same criteria, and refrain from physically tilting your head or rotating your display."*
- **Shadow swap:** Participants were instructed: *"Rate the naturalness of the shadow shape formed on the floor surface on a 5-point scale. If the shadow shape appears consistent with the object shape, select 5 (Natural). If the shadow shape appears inconsistent with the object shape, select 1–4 according to the perceived unnaturalness."*
- **Reflection swap:** Participants were instructed: *"Rate the naturalness of the reflection on the floor surface on a 5-point scale. If the appearance of the reflected image matches the physical object, select 5 (Natural); if it does not match, select 1–4 according to the degree of unnaturalness."*
- **Light-direction swap:** Participants were instructed: *"In an image depicting three objects illuminated from above, rate the naturalness of the lighting on the central object on a 5-point scale, focusing on whether the lighting direction on the central object is consistent with the surrounding objects and the floor shadows. If the lighting direction is mutually consistent, select 5 (Natural); if it appears contradictory, select 1–4 according to the degree of unnaturalness."*

B.3 EXPERIMENTAL DESIGN AND TRIAL STRUCTURE

Thatcher task. The 140 base face scenes (140×2 orientations $= 280$ pairs; 560 images) were divided into four disjoint sets of 35 (A–D). The four observer groups viewed A upright and B inverted, B upright and C inverted, C upright and D inverted, or D upright and A inverted. Each participant therefore viewed 35 upright faces and 35 different inverted faces, with each identity presented in only one orientation. Both original and modified versions of each scene were evaluated by the same participant, providing within-subject paired perceptual differences $h_i = R(x_i^{\text{orig}}) - R(x_i^{\text{mod}})$. Each main block contained 140 trials (35 upright faces plus 35 different inverted faces, each shown in

original and modified versions) and 6 repeated trials for reliability checking (146 trials total). In data processing, repeated trials were deduplicated by retaining the first presentation.

Illumination task. The three illumination interventions (Shadow, Reflection, Light direction; 96 scenes each, 288 pairs total) were evaluated in a between-subjects design: each participant evaluated only one intervention type. Within each intervention, the 96 scenes were randomly divided into four disjoint subsets of 24. Each of four observer groups evaluated a different combination of three subsets (72 scenes), so each scene was evaluated by three of the four groups. Both original and modified versions of each assigned scene were evaluated within-subject. Each main block contained 144 trials (72 scenes $\times$ 2 versions) and 6 repeated trials (150 trials total; first presentation retained).

Procedure. Each session progressed through: full-screen switch $\rightarrow$ illustrated instructions $\rightarrow$ a practice block (10 trials for Thatcher; condition-specific practice for Illumination) $\rightarrow$ an option to re-read instructions $\rightarrow$ the main block with trials presented in randomized order $\rightarrow$ data submission.

B.4 Participant demographics, screening, and response counts

Participants were adults aged 20–49 years, with balanced representation across age groups (20s, 30s, and 40s) and sexes. A total of 162 participant submissions were collected for the Thatcher task and 167 for the Illumination task. Since two Illumination submissions came from duplicate participant identities, we retained each participant's first submission while discarding the latter, leaving 165 participants before response-consistency screening. Although the experiment included 6 repeated trials originally intended for intra-subject test–retest screening, strict consistency filtering on these few trials proved overly conservative due to normal perceptual rating jitter on the 5-point scale, excluding an unnecessarily large fraction of otherwise attentive participants. We therefore adopted a global response-consistency screening criterion. Repeated trials were deduplicated by retaining the first presentation, and for each participant, we computed the Pearson correlation r between their ratings across unique images and the corresponding image-wise mean ratings of all other participants who rated those images.

We estimated a Gaussian kernel density for each task using Silverman's rule-of-thumb bandwidth (Silverman, 1986), $b = 0.9 \min(s, \mathrm{IQR}/1.34) n^{-1/5}$, where s, IQR, and n are the sample standard deviation, interquartile range, and number of finite correlations. We selected the deepest trough relative to its flanking peaks (Figure S9), yielding cutoffs of $r = .2481$ for Thatcher and $r = .3236$ for Illumination. Retaining participants who met or exceeded these thresholds left 143 participants for Thatcher (88.3%) and 105 for Illumination (63.6%).

After screening, image items received the following numbers of independent human ratings:

- **Thatcher task:** Mean 35.8 responses per item (range: 29–44 responses; upright mean: 35.8, inverted mean: 35.8).
- **Illumination task:** Overall mean 26.2 responses per item (range: 9–43). By condition, Shadow swap averaged 40.5 responses (range: 38–43; 54 participants retained), Reflection swap averaged 26.2 responses (range: 22–29; 35 participants), and Light-direction swap averaged 12.0 responses (range: 9–14; 16 participants).

The variation in sample size across Illumination conditions stemmed primarily from differences in exclusion rates during screening: before screening, the Shadow, Reflection, and Light-direction conditions enrolled 69, 44, and 52 participants, of whom 15 (21.7%), 9 (20.5%), and 36 (69.2%), respectively, were excluded. In participant-disjoint split-half analyses (10,000 group-stratified random splits), split-half reliability in the Light-direction condition was $r = .144$ among excluded participants, compared with $r = .668$ among retained participants (and .960 for Shadow, .932 for Reflection). The excluded Light-direction group showed little average separation between original and modified ratings: the mean original-minus-modified difference across the 96 pairs was 0.051 points, compared with 0.996 points in the retained group.

Filtering robustness. To verify that our findings do not depend on the specific screening threshold, we recomputed all item-level aggregates without any participant exclusions. The unfiltered item aggregates correlated closely with the screened aggregates ($r = .986$ for Thatcher, $r = .942$ for

Illumination), and the cross-model sensitivity–alignment association was preserved ($r = .663$ with and without screening on Thatcher; .821 with screening and .847 without screening on Illumination). Because Light direction had the highest exclusion rate, we checked whether screening affected model–human correspondence in this condition. Across the 96 image pairs, the mean original–modified rating differences calculated from all 52 participants closely tracked those from the 16 retained participants ($r = .891$). Model–human correlations were also similar (median across all 25 generative models: .277 with screening and .255 without).

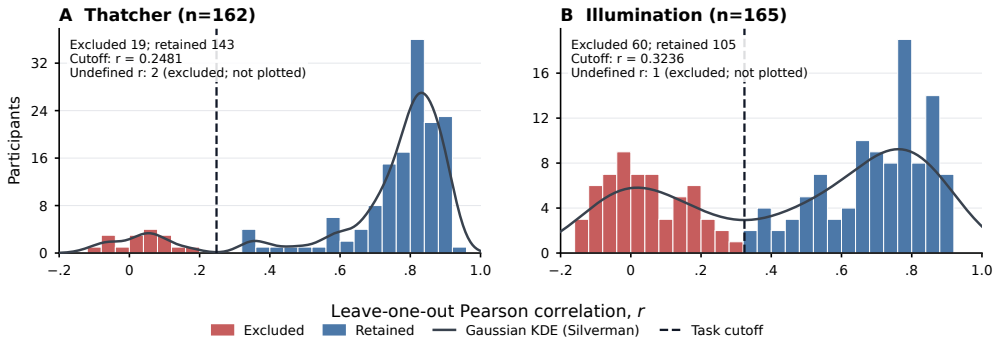

Figure S9: **Participant response consistency and task-specific screening cutoffs.** Distributions of leave-one-out Pearson correlations r for (A) Thatcher ($n = 162$) and (B) Illumination ($n = 165$), after duplicate submissions and repeated presentations were reduced by retaining the first occurrence. Each participant's ratings are correlated with the image-wise means of the other participants. Red and blue bars denote excluded and retained participants, respectively; histogram bins have width .04. Solid curves show Gaussian kernel density estimates with Silverman's bandwidths ($b = .03836$ and .11157), scaled to histogram counts. Dashed lines mark the task-specific cutoffs ($r = .2481$ and .3236). Undefined correlations (2 for Thatcher, 1 for Illumination) are excluded from the plots and counted among excluded participants.

C Model Details and Evaluation Specifications

This section summarizes the common loss probing setup and specifications for all 25 evaluated generative models (Table S1). Model-specific schedulers, identifiers, and environment configurations are archived with the accompanying code and configuration files.

C.1 Generative Loss Probing

For each generative model m and test image x, the model-native loss $L_m(x)$ is computed by averaging its squared prediction error across a fixed grid of timesteps T_m and noise realizations E_m:

$$L_m(x) = \frac{1}{|T_m||E_m|} \sum_{t \in T_m} \sum_{\epsilon \in E_m} \ell_m(x, t, \epsilon). \tag{1}$$

Here, ℓ_m is the mean squared error in the parameterization of the model's denoising objective (noise, velocity, or flow), averaged over channel, spatial, and, where applicable, temporal dimensions. For latent-space models, losses are evaluated on encoded inputs and do not include an image-reconstruction term. JiT and PixelGen (Li & He, 2026; Ma et al., 2026) predict clean images and compute their denoising losses in velocity space. We follow this definition, using the residual $(\hat{x} - x)/\max(1 - t, 0.05)$, where $\hat{x}$ is the predicted clean image, t is the clean-image interpolation coefficient, and the denominator floor follows the original implementations.

Timestep grids (T_m). Image models were evaluated at 100 scheduler-defined timesteps and video models at 50. The grids followed each model's implemented variance-preserving or shifted-flow schedule and were averaged with equal weight; their numerical timestep coordinates are model-specific and not directly comparable across architectures.

Noise realizations (E_m). Each model evaluation used a fixed bank of 20 standard Gaussian noise tensors. This bank was instantiated once and shared across all stimuli and both members of each pair, isolating content-dependent differences under matched noise draws.

Conditioning and guidance. Unconditional models used their native null-class or unconditional branch. Text-conditioned models used an empty or neutral prompt in a single forward pass without classifier-free guidance (CFG) branch mixing. The embedded guidance scale was set to 1.0 for both FLUX.1 dev and HunyuanVideo.

Image preprocessing. All 16 image models received RGB inputs resized to 512×512 pixels before model-specific normalization or latent encoding.

Video preprocessing. Each source RGB image was resized by bilinear interpolation to fit the model's input resolution (Table S1), preserving aspect ratio, then center-padded with a constant value of .5. The padded frame was repeated to form a static clip. CogVideoX1.5 uses 21 repeats to yield six latent frames, satisfying its two-frame temporal patch size; the other eight models use 17.

Compute environment and numerical precision. Model probing was executed within isolated container environments on NVIDIA RTX A6000 (48 GB) and A100 (40 GB) GPUs, using native `bfloat16`, `float16`, or `float32` precision compatible with each model. Full software dependencies and configuration files are provided in the supplementary material.

Table S1: **Evaluated open generative models.** Specifications of the 25 generative models across 11 families, grouped by architectural domain: pixel-space image models (circles in figures), latent-space image models (squares), and spatiotemporal video models (triangles). Superscripts ^{T} and ^{I} mark the top five generative models selected by overall human alignment in Thatcher and Illumination, respectively; these selections are held fixed across conditions (Section I).

Family	Model	Identifier / Source Repository	Architecture	Evaluated Resolution	Steps ($\lvert T_m \rvert$)	Noise ($\lvert E_m \rvert$)
Pixel-space image models						
JiT	JiT-B/32	`LTH14/JiT::jit-b-32/checkpoint-last.pth`	Pixel DiT (ImageNet)	512×512	100	20
JiT	JiT-L/32	`LTH14/JiT::jit-l-32/checkpoint-last.pth`	Pixel DiT (ImageNet)	512×512	100	20
JiT	JiT-H/32	`LTH14/JiT::jit-h-32/checkpoint-last.pth`	Pixel DiT (ImageNet)	512×512	100	20
PixelGen	PixelGen 512	`zehongma/PixelGen::PixelGen_XXL_T2I.ckpt`	Pixel DiT / Flow	512×512	100	20
HiDream	HiDream-O1-Image	`HiDream-ai/HiDream-O1-Image`	Pixel-space UiT / Flow	512×512	100	20
Latent-space image models						
Stable Diffusion	SD v1.5^{T}	`sd-legacy/stable-diffusion-v1-5`	Latent U-Net (DDPM)	512×512	100	20
Stable Diffusion	SDXL Base 1.0	`stabilityai/stable-diffusion-xl-base-1.0`	Latent U-Net (EDM)	512×512	100	20
Stable Diffusion	SD 3 Medium	`stabilityai/stable-diffusion-3-medium-diffusers`	Latent MMDiT (Flow)	512×512	100	20
FLUX.1	FLUX.1 dev	`black-forest-labs/FLUX.1-dev`	Latent MMDiT (Flow)	512×512	100	20
FLUX.1	FLUX.1 schnell^{T}	`black-forest-labs/FLUX.1-schnell`	Latent MMDiT (Distilled)	512×512	100	20
FLUX.2	FLUX.2 klein base 4B	`black-forest-labs/FLUX.2-klein-base-4B`	Latent MMDiT (Flow)	512×512	100	20
FLUX.2	FLUX.2 klein 4B^{T}	`black-forest-labs/FLUX.2-klein-4B`	Latent MMDiT (Flow)	512×512	100	20
FLUX.2	FLUX.2 klein base 9B	`black-forest-labs/FLUX.2-klein-base-9B`	Latent MMDiT (Flow)	512×512	100	20
FLUX.2	FLUX.2 klein 9B^{T}	`black-forest-labs/FLUX.2-klein-9B`	Latent MMDiT (Flow)	512×512	100	20
Qwen-Image	Qwen-Image^{I}	`Qwen/Qwen-Image`	Latent MMDiT (Flow)	512×512	100	20
Qwen-Image	Qwen-Image-2512^{I}	`Qwen/Qwen-Image-2512`	Latent MMDiT (Flow)	512×512	100	20
Spatiotemporal video models (17 frames; CogVideoX1.5: 21)						
CogVideoX	CogVideoX-2B	`THUDM/CogVideoX-2b`	3D VAE Latent DiT	720×480	50	20
CogVideoX	CogVideoX-5B	`THUDM/CogVideoX-5b`	3D VAE Latent DiT	720×480	50	20
CogVideoX	CogVideoX1.5-5B	`THUDM/CogVideoX1.5-5B`	3D VAE Latent DiT	1360×768	50	20
LTX-Video	LTX-Video 2B v0.9.0	`Lightricks/LTX-Video`	3D VAE Latent DiT	768×512	50	20
LTX-Video	LTX-Video 13B dev^{T}	`Lightricks/LTX-Video-0.9.7-dev`	3D VAE Latent DiT	768×512	50	20
Wan	Wan2.1 T2V 1.3B	`Wan-AI/Wan2.1-T2V-1.3B-Diffusers`	3D VAE Latent DiT	832×480	50	20
Wan	Wan2.1 T2V 14B^{I}	`Wan-AI/Wan2.1-T2V-14B-Diffusers`	3D VAE Latent DiT	832×480	50	20
Wan	Wan2.2 T2V A14B^{I}	`Wan-AI/Wan2.2-T2V-A14B-Diffusers`	3D VAE Latent DiT (MoE)	832×480	50	20
HunyuanVideo	HunyuanVideo^{I}	`hunyuanvideo-community/HunyuanVideo`	3D VAE Latent DiT	960×544	50	20

HiDream-O1 in figure labels abbreviates HiDream-O1-Image. LTX-Video 2B uses the initial v0.9.0 checkpoint in Diffusers format.

D NON-GENERATIVE BASELINE DETAILS AND SCORING METHODS

To benchmark generative loss against non-generative alternatives, we evaluate a heterogeneous pool of 12 frozen image encoders alongside two prespecified five-metric IQA batteries (full-reference and no-reference; Table S3). All model weights and scoring protocols are held fixed without task-specific fine-tuning.

Encoders were selected by predefined criteria independent of observed human alignment to represent standard anchors (ResNet-50, CLIP, and DINOv2), diverse pretraining objectives, varying model scales (0.024B–8.145B parameters), and within-family scaling pairs (e.g., Qwen3-VL-Embedding 2B vs. 8B, PE-Core B/16 vs. G/14). The resulting pool consists of 12 frozen encoders spanning supervised classification (ResNet-50; He et al., 2016), self-supervised visual representation learning (DINOv2, DINOv3; Oquab et al., 2024; Sim'eoni et al., 2026), and contrastive image–language alignment (CLIP, SigLIP 2, PE-Core, Jina CLIP v2; Radford et al., 2021; Tschannen et al., 2025; Bolya et al.,

2025; Koukounas et al., 2024). Qwen3-VL-Embedding provides a multimodal embedding baseline trained for representation extraction, using the representation of the last valid input token (Li et al., 2026).

D.1 Frozen encoder distances

For model m and block ℓ, let $\phi_{m\ell}(x)$ denote the feature obtained by applying a fixed model-specific readout to that block's output. For global image representations, the pair score is the unsigned cosine distance between independently extracted representations of the original and modified images:

$$d_{\text{enc}}(x_i^{\text{orig}}, x_i^{\text{mod}}) = 1 - \frac{\phi(x_i^{\text{orig}})^\top \phi(x_i^{\text{mod}})}{\|\phi(x_i^{\text{orig}})\|_2 \|\phi(x_i^{\text{mod}})\|_2}. \tag{2}$$

Here $\phi = \phi_{m\ell}$ for the candidate being evaluated. No classifier or readout is fitted to human judgments.

Because the 12 encoders differ in their internal architectures (spanning convolutional feature maps, Vision Transformers with or without class tokens, and multimodal language decoders), extracting comparable representations $\phi_{m\ell}(x)$ requires architecture-specific readout rules. We hold these readout rules fixed across all blocks and tasks. Below, we specify how representations are extracted across three structural groups: models using direct global pooling or unprojected class tokens (ResNet-50, CLIP, and PE-Core); models evaluated via spatially matched patch-token distances (DINOv2 and DINOv3); and models using terminal pooling, projection, or normalization transformations (SigLIP 2, Jina CLIP v2, and Qwen3-VL-Embedding).

ResNet, CLIP, and PE-Core. ResNet-50 uses spatial global average pooling after each of its 16 bottleneck blocks, without the classification head. CLIP uses the block's unprojected `[CLS]` token, without the terminal LayerNorm or projection. This choice follows the default unnormalized intermediate-feature convention in OpenCLIP, while retaining the original OpenAI CLIP checkpoint. PE-Core uses the spatial mean of block tokens after excluding prefix tokens, without terminal normalization or attention pooling. The PE feature API exposes unnormalized intermediate tokens; their spatial averaging is our fixed aggregation choice. Pooling rules therefore follow the native token structure of each architecture.

DINOv2/v3 spatial distances. For DINOv2 and DINOv3, we apply the pretrained terminal LayerNorm to each candidate block, following the normalized intermediate-feature convention. We then exclude the leading `[CLS]` token and, for DINOv3, four register tokens. The score is the mean one-minus-cosine distance between corresponding spatial patch tokens. Tokens are not averaged before computing cosine distance, and no self-supervised projection head is used. This preserves a spatially matched comparison under the controlled original–modified pairing.

SigLIP, Jina, and Qwen output transformations. SigLIP 2 applies its pretrained terminal LayerNorm and learned attention pooling to each block. Jina CLIP v2 applies its terminal normalization, `[CLS]` extraction, and projection. Qwen3-VL-Embedding processes images through its full visual tower into a multimodal language decoder (28 blocks for 2B, 36 for 8B). Following its official representation extraction protocol, we apply terminal RMSNorm and ℓ_2 normalization to the representation of the last valid input token at each decoder block. Both Qwen models use the fixed default instruction "Represent the user's input." without task-specific prompting. These transformations are applied directly to intermediate blocks without retraining.

Candidate layer evaluation. To identify the strongest representations achievable by each architecture, we evaluated pair distances across all 314 intermediate backbone blocks alongside each model's default final output (e.g., CLIP's projected embedding or PE-Core's attention-pooled vector). For each task, the representative layer reported in Table S2 was selected by maximizing the signed Pearson correlation with human judgment differences across all stimulus pairs. The downstream protocol for incorporating these selected layers and top-performing models into partial correlation analyses is described in Appendix E.3.

Table S2: **Selected encoder layers.** Entries indicate the selected layer index and total model depth for each task. Entries marked "(output)" use the model's default final embedding rather than intermediate block readouts (Section D.1). For Qwen, indices denote language-decoder blocks. Selected layers remain fixed across all condition-specific comparisons within each task.

Encoder	Thatcher	Illumination
ResNet-50	7/16	16/16 (output)
CLIP ViT-B/16	12/12	9/12
DINOv2-Base	2/12	12/12 (output)
DINOv2-g/14	10/40	8/40
DINOv3-H+/16	32/32 (output)	1/32
DINOv3-7B/16	36/40	5/40
SigLIP 2 Base/16	9/12	7/12
PE-Core B/16	12/12 (output)	11/12
PE-Core G/14	50/50 (output)	33/50
Jina CLIP v2	23/24	22/24
Qwen3-VL-Embedding 2B	22/28	28/28 (output)
Qwen3-VL-Embedding 8B	31/36	35/36

D.2 Full-reference IQA (FR-IQA)

Full-reference metrics evaluate the distortion of x_i^{mod} relative to the pristine original reference x_i^{orig}. To align all FR-IQA metrics along a consistent *degradation / distance* axis where larger values indicate greater unnaturalness or perceptual difference:

$$u_{\text{FR}}(x_i^{\text{orig}}, x_i^{\text{mod}}) = \begin{cases} -\text{Metric}(x_i^{\text{mod}}, x_i^{\text{orig}}) & \text{if higher is better (PSNR, SSIM, VIF),} \\ \text{Metric}(x_i^{\text{mod}}, x_i^{\text{orig}}) & \text{if lower is better (LPIPS-VGG, DISTS).} \end{cases} \tag{3}$$

Specifically, PSNR is computed from pixel-level mean squared error; SSIM evaluates local luminance, contrast, and structure over sliding windows; VIF models visual information fidelity in the wavelet domain; LPIPS-VGG computes ℓ_2-distance across five intermediate VGG-16 feature activation layers weighted by BAPPS perceptual judgments; and DISTS evaluates multi-scale structural and texture similarities using spatial deep features.

D.3 No-reference IQA (NR-IQA)

No-reference metrics score each image independently without access to a reference image: $q(x)$. To measure directional quality changes on original–modified pairs, scores are differenced strictly according to each metric's published valence:

$$u_{\text{NR}}(x_i^{\text{orig}}, x_i^{\text{mod}}) = \begin{cases} q(x_i^{\text{orig}}) - q(x_i^{\text{mod}}) & \text{if higher is better (HyperIQA, MUSIQ, LIQE),} \\ q(x_i^{\text{mod}}) - q(x_i^{\text{orig}}) & \text{if lower is better (NIQE, BRISQUE).} \end{cases} \tag{4}$$

Under this definition, positive values ($u_{\text{NR}} > 0$) consistently signify that the original image is evaluated as higher quality than the modified image. The battery spans NIQE, an opinion-unaware natural-scene-statistics (NSS) model fitted to pristine luminance images; BRISQUE, spatial NSS features paired with an opinion-trained support vector regressor; HyperIQA, a CNN hypernetwork predicting content-adaptive quality weights; MUSIQ, a multi-scale Vision Transformer trained on KonIQ-10k; and LIQE, a multimodal quality evaluator using CLIP semantic prompts trained on KonIQ.

D.4 Preprocessing and execution environment

All input images were resized to 512×512 pixels without cropping, followed by model-specific channel normalization. IQA metrics used their default processors from PyIQA (v0.1.9) (Chen & Mo, 2022). Probing was executed on NVIDIA RTX A6000 GPUs, with dependencies and configurations provided in the supplementary material.

Table S3: **Evaluated non-generative baseline suite.** Specifications of the 12 frozen vision encoders alongside five full-reference (FR-IQA) and five no-reference (NR-IQA) metrics. Parameter counts (B) report total network parameters in billions. Superscripts ^{T} and ^{I} mark the top five encoders selected by overall human alignment on Thatcher and Illumination, respectively; these selections are held fixed across conditions (Section I).

Pool	Predictor	Identifier / checkpoint	Training objective	Params. (B)	Readout / score
Encoder	ResNet-50 (ImageNet-1K, torchvision V2)	`torchvision::IMAGENET1K_V2`	ImageNet supervision	.024	Block spatial mean; cosine distance
Encoder	CLIP ViT-B/16$^{\mathrm{T,I}}$	`openai/clip-vit-base-patch16`	Image–text contrastive	.086	Unprojected block CLS; cosine distance
Encoder	DINOv2-Base	`facebook/dinov2-base`	Vision-only self-supervision	.087	Corresponding patch-token distance
Encoder	SigLIP 2 Base/16^{I} (224)	`google/siglip2-base-patch16-224`	Image–text sigmoid contrastive	.093	LayerNorm, attention pooling; cosine distance
Encoder	PE-Core B/16^{I} (224)	`facebook/PE-Core-B16-224`	Vision–language representation	.094	Block spatial mean; cosine distance
Encoder	DINOv2-g/14 spatial	`facebook/dinov2-giant`	Vision-only self-supervision	1.136	Corresponding patch-token distance
Encoder	PE-Core G/14$^{\mathrm{T,I}}$ (448)	`timm/vit_pe_core_gigantic_patch14_448.fb`	Vision–language representation	1.882	Block spatial mean; cosine distance
Encoder	DINOv3-H+/16 spatial^{T}	`facebook/dinov3-vith16plus-pretrain-lvd1689m`	Vision-only self-supervision	.840	Corresponding patch-token distance
Encoder	DINOv3-7B/16 spatial^{T}	`facebook/dinov3-vit7b16-pretrain-lvd1689m`	Vision-only self-supervision	6.716	Corresponding patch-token distance
Encoder	Qwen3-VL-Embedding 2B	`Qwen/Qwen3-VL-Embedding-2B`	Multimodal embedding	2.128	RMSNorm, last valid token; cosine distance
Encoder	Qwen3-VL-Embedding 8B^{T}	`Qwen/Qwen3-VL-Embedding-8B`	Multimodal embedding	8.145	RMSNorm, last valid token; cosine distance
Encoder	Jina CLIP v2^{I}	`jinaai/jina-clip-v2`	Multilingual image–text contrastive	.304	Terminal norm, CLS projection; cosine distance
Full-reference IQA					
FR-IQA	PSNR	`pyiqa::psnr`	Pixel error	—	$-\mathrm{PSNR}(x^{\mathrm{mod}}, x^{\mathrm{orig}})$
FR-IQA	SSIM	`pyiqa::ssim`	Structural similarity	—	$-\mathrm{SSIM}(x^{\mathrm{mod}}, x^{\mathrm{orig}})$
FR-IQA	VIF	`pyiqa::vif`	Visual information fidelity	—	$-\mathrm{VIF}(x^{\mathrm{mod}}, x^{\mathrm{orig}})$
FR-IQA	LPIPS-VGG	`pyiqa::lpips-vgg`	Learned perceptual distance	—	$\mathrm{LPIPS}(x^{\mathrm{mod}}, x^{\mathrm{orig}})$
FR-IQA	DISTS	`pyiqa::dists`	Structure and texture similarity	—	$\mathrm{DISTS}(x^{\mathrm{mod}}, x^{\mathrm{orig}})$
No-reference IQA					
NR-IQA	NIQE	`pyiqa::niqe`	Natural scene statistics	—	$q(x^{\mathrm{mod}}) - q(x^{\mathrm{orig}})$
NR-IQA	BRISQUE	`pyiqa::brisque`	Natural scene statistics + SVR	—	$q(x^{\mathrm{mod}}) - q(x^{\mathrm{orig}})$
NR-IQA	HyperIQA	`pyiqa::hyperiqa`	CNN hypernetwork	—	$q(x^{\mathrm{orig}}) - q(x^{\mathrm{mod}})$
NR-IQA	MUSIQ	`pyiqa::musiq(koniq10k)`	Multi-scale Transformer	—	$q(x^{\mathrm{orig}}) - q(x^{\mathrm{mod}})$
NR-IQA	LIQE	`pyiqa::liqe(koniq)`	Language-prompted IQA	—	$q(x^{\mathrm{orig}}) - q(x^{\mathrm{mod}})$

E STATISTICAL PROCEDURES AND ANALYSIS DETAILS

This section provides methodological details for the statistical inference, reliability estimation, and baseline partial correlation analysis reported in the main text.

E.1 BOOTSTRAP RESAMPLING AND STATISTICAL UNCERTAINTY

To account for sampling uncertainty across both observers and stimuli, primary human-dependent estimates (such as model–human alignment and human directional sensitivity) are evaluated via a nonparametric two-factor bootstrap (10,000 draws with fixed random seeds), independently resampling participants and base scenes with replacement. Participants are sampled within each stimulus-assignment group at their original sample sizes. Stimuli are sampled at the level of base scenes: Thatcher stimuli were resampled across the 140 face identities (keeping upright and inverted pairs together); Illumination stimuli were resampled at the level of the 48 base configurations within each operation, keeping horizontally mirrored variants together. Generative-model sensitivities do not depend on human ratings and are evaluated using base-scene resampling alone. In each iteration, condition means and the task-wide standard deviation denominator ($s_{m,\text{task}}$) are recomputed dynamically, propagating scaling uncertainty into the resulting estimates. Identical bootstrap draws are shared across all models, conditions, and paired comparisons within each task. Pointwise 95% confidence intervals correspond to the empirical 2.5th and 97.5th percentiles across draws, summarizing sampling variability across observers and stimuli.

Human factor-profile intervals. The error bars for human profile means in Figure 1B and Appendix J evaluate observer sampling variability conditional on the fixed stimulus set. These pointwise 95% intervals are constructed from 10,000 participant-bootstrap iterations within each stimulus-assignment group, recomputing image-wise means and condition-scaled effects within each draw.

Cross-model associations. To evaluate uncertainty in cross-model sensitivity–alignment correlations while accounting for dependence among related models, we combine the participant-and-scene resampling above with a cluster bootstrap over 11 model families (CogVideoX, FLUX.1, FLUX.2, HiDream, HunyuanVideo, JiT, LTX-Video, PixelGen, Qwen-Image, Stable Diffusion, and Wan). In each of 10,000 iterations, 11 families are sampled with replacement, retaining all models within each selected family to recompute the cross-model correlation. Resampled family draws are held consistent across tasks and conditions. Reported 95% confidence intervals correspond to the pointwise 2.5th–97.5th percentiles across iterations, summarizing joint sampling variability across observers, stimuli, and model families.

E.2 HUMAN SPLIT-HALF RELIABILITY ESTIMATION

To establish an empirical reference for human rating consistency, we evaluate split-half reliability across 10,000 random participant partitions. Within each stimulus-assignment group, participants are randomly divided into two disjoint halves (A and B) of equal (or near-equal) size, holding partitions consistent across experimental conditions. For each partition, item-level mean unnaturalness effects ($h_{A,i}$ and $h_{B,i}$) are computed independently for each half, and split-half reliability is evaluated as their Pearson correlation:

$$r_{\text{half}} = \text{corr}_i(h_{A,i}, h_{B,i}). \tag{5}$$

In Figures 3 and 5 (and their condition breakdowns, Figures S13 and S14), reference lines and markers indicate the mean split-half correlation across partitions on the complete stimulus set. To reflect variability from both rater partitioning and stimulus sampling, each partition is paired with a synchronized base-scene bootstrap draw, with the 2.5th and 97.5th percentiles defining the 95% resampling intervals (shown as shaded bands and error bars).

E.3 BIDIRECTIONAL PARTIAL CORRELATION ANALYSIS

To test whether Generative scores retain human-aligned variation beyond strong frozen encoders and quality metrics, we compare the 25 Generative scores with the 12-encoder pool and the two prespecified five-metric IQA batteries described in Section 3.4 and Appendix D.

Model and layer selection. To compare generative models against the strongest non-generative baselines, top-performing models are selected in two steps. First, for each of the 12 frozen encoders, we identify the layer that achieves the highest Pearson correlation with human unnaturalness ratings across the full task (Table S2). Second, from the 25 generative models and the 12 selected-layer encoders, we rank models by their correlation with human ratings and retain the top five from each pool. All five metrics in each IQA battery are retained without selection. These selected top-five models and encoder layers are held constant across all analyses within each task, including the condition breakdowns and factor profiles (Appendix I), ensuring that differences across conditions reflect stimulus properties rather than shifts in model composition.

Bidirectional partial correlations. For every selected Generative–comparator pair (G, X), we compute

$$r(h, G \mid X) = \frac{r_{hG} - r_{hX} r_{GX}}{\sqrt{(1 - r_{hX}^2)(1 - r_{GX}^2)}}, \qquad r(h, X \mid G) = \frac{r_{hX} - r_{hG} r_{GX}}{\sqrt{(1 - r_{hG}^2)(1 - r_{GX}^2)}}. \tag{6}$$

The two directional summaries are the medians of the 25 values in the selected 5×5 grid.

Accounting for selection uncertainty in bootstrap intervals. To include participant, stimulus, and selection variation, confidence intervals use the same 10,000 participant-and-scene bootstrap draws. On each draw, the two-step selection process is repeated on the resampled full-task data: identifying the best-performing layer for each encoder, and then ranking models to select the top five generative models and encoders. These selections are applied consistently across the overall task and condition breakdowns for that draw. For each comparison, 95% confidence intervals for the directional partial correlations are obtained from the 2.5th and 97.5th percentiles across the bootstrap distribution.

F Diagnostic controls on native generative loss

Generative model loss is heavily sensitive to low-level image complexity and high-frequency statistics (Nalisnick et al., 2019; Serrà et al., 2020). Consequently, paired loss differences ($L_{\text{modified}} - L_{\text{original}}$) fail to track perceptual naturalness whenever an intervention alters these low-level properties, and unconditioned standalone losses fail to reflect human ratings across differing visual compositions. To evaluate the necessity of content-preserving relational manipulations and paired score differencing, we conducted two diagnostic evaluations: generic low-level distortions alongside element deletion, and standalone single-image loss evaluations.

F.1 Generic distortions on CSIQ and element deletion

To test these limitations, we evaluated paired model loss under two scenarios: generic low-level distortions on the CSIQ benchmark (Larson & Chandler, 2010) and element deletion in synthetic 3D scenes (Figure S10).

Generic distortions on CSIQ. First, we tested whether loss differences relative to the reference image ($L_{\text{distorted}} - L_{\text{reference}}$) track human quality judgments across standard corruptions using the CSIQ benchmark (Larson & Chandler, 2010) (Figure S10A). We restricted this evaluation to pixel-space diffusion models (three JiT variants, PixelGen 512, and HiDream-O1-Image) to avoid confounding generative priors with VAE compression characteristics. Evaluating loss–DMOS correlations within each distortion type reveals a severe statistical bias (Figure S10B). For distortions that add high-frequency variance (AWGN and pink noise), model loss correlates positively with perceived degradation. In contrast, for distortions that smooth textures or attenuate contrast (blur, JPEG, JPEG 2000, and contrast reduction), the correlation systematically reverses to negative. Because denoising objectives naturally penalize high-frequency variance while favoring smoother images, paired loss differences cannot overcome unconstrained shifts in low-level statistics, rendering native loss ineffective as a general-purpose quality metric.

Element deletion versus relational relocation. Second, we examined whether altering image content undermines the evaluation of physical consistency. Using 96 rendered 3D scenes, we compared model responses under two shadow manipulations (Figure S10C): (1) *one-shadow deletion*,

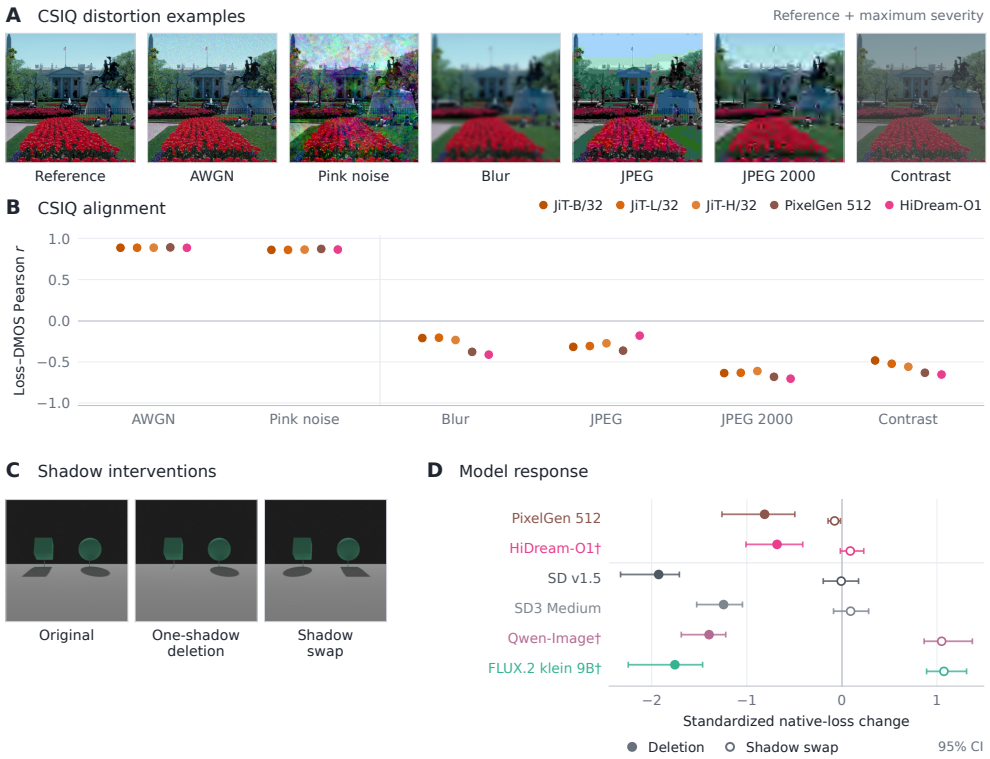

Figure S10: **Impact of low-level image statistics on native generative loss.** (A) Illustrative CSIQ reference image alongside corrupted examples at maximum severity across six distortion types. (B) Paired loss difference ($L_{\text{distorted}} - L_{\text{reference}}$)–DMOS Pearson correlations within each CSIQ distortion type across five pixel-space diffusion models. (C) Example synthetic scenes: original baseline, one-shadow deletion, and shadow swap (relocation). (D) Mean loss changes ($L_{\text{modified}} - L_{\text{original}}$), divided by their sample standard deviation over the 96 pairs of each manipulation, for one-shadow deletion (filled circles) versus shadow swap (open circles). Error bars denote percentile 95% intervals from 10,000 resamples of the 48 base scenes.

which removes a cast shadow entirely, introducing a physical anomaly while reducing overall image complexity; and (2) *shadow swap* (our main paradigm), which relocates shadows between objects, creating physical inconsistency while preserving low-level image statistics.

Standardized loss differences (defined for this control as the mean paired loss difference divided by the sample standard deviation over the 96 pairs) show distinct patterns between the two manipulations (Figure S10D). For one-shadow deletion, loss changes were uniformly *negative* across all models: although removing a shadow makes the scene physically unnatural, the reduced image complexity dominates the loss landscape, causing the model to evaluate the corrupted scene as more probable. In contrast, under shadow swap, preserving low-level statistics mitigates this negative artifact, allowing more capable architectures to exhibit positive loss penalties that reflect sensitivity to the physical inconsistency.

F.2 SINGLE-IMAGE ABSOLUTE SCORE ANALYSIS

Raw single-image model losses ($L_m(x)$ evaluated without reference-image subtraction) do not show consistently positive alignment with human unnaturalness judgments ($a(x) = 5 - R(x)$) across image variants and tasks (Figure S11). Across 280 Thatcher and 288 Illumination pairs, we correlated each original- or modified-image loss with the rating of that same image. The 25-model median Pearson correlations were .236 (original) and $-.090$ (modified) on Thatcher faces, and $-.145$ (original) and $-.203$ (modified) on Illumination scenes. Paired score differences ($u_{m,i} = L_m(x_i^{\text{mod}}) - L_m(x_i^{\text{orig}})$) correlated positively with human paired differences ($h_i = R(x_i^{\text{orig}}) - R(x_i^{\text{mod}})$) for all 25 models in both tasks, with median correlations of .747 on Thatcher and .334 on Illumination.

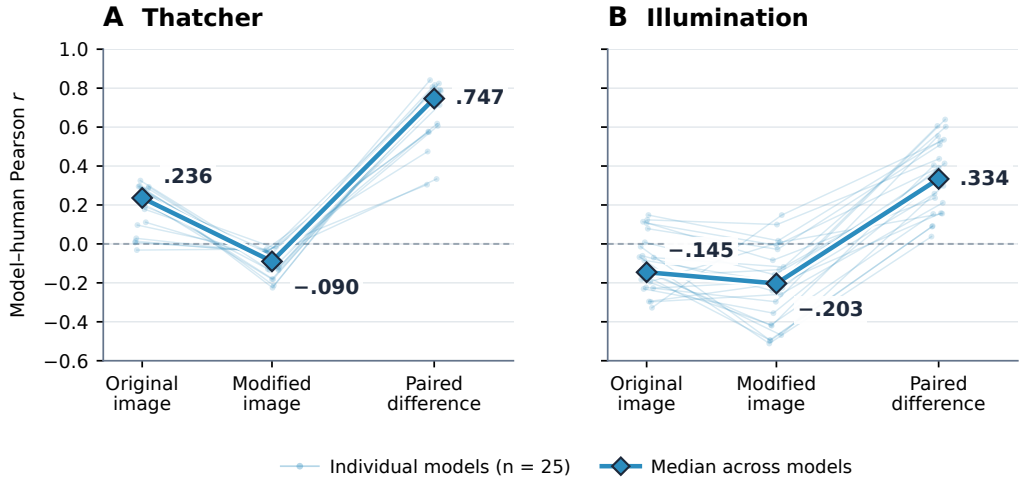

Figure S11: **Single-image losses and paired loss changes in relation to human judgments.** Item-level Pearson correlations for all 25 generative models, pooled across conditions within (A) Thatcher and (B) Illumination. Original- and modified-image losses are correlated with the corresponding image's mean unnaturalness, $5 - R(x)$; paired loss changes are correlated with $R(x^{\text{orig}}) - R(x^{\text{mod}})$. Faint lines and small points show individual models, with each line connecting estimates from the same model. Thick lines, diamonds, and numeric labels indicate the median across all 25 models.

G MODEL-WISE SENSITIVITY AND HUMAN ALIGNMENT

Figure S12 shows directional sensitivity and human alignment for all 25 generative models across the five stimulus conditions. Appendix H shows the relationship between these two measures as condition-specific scatter plots.

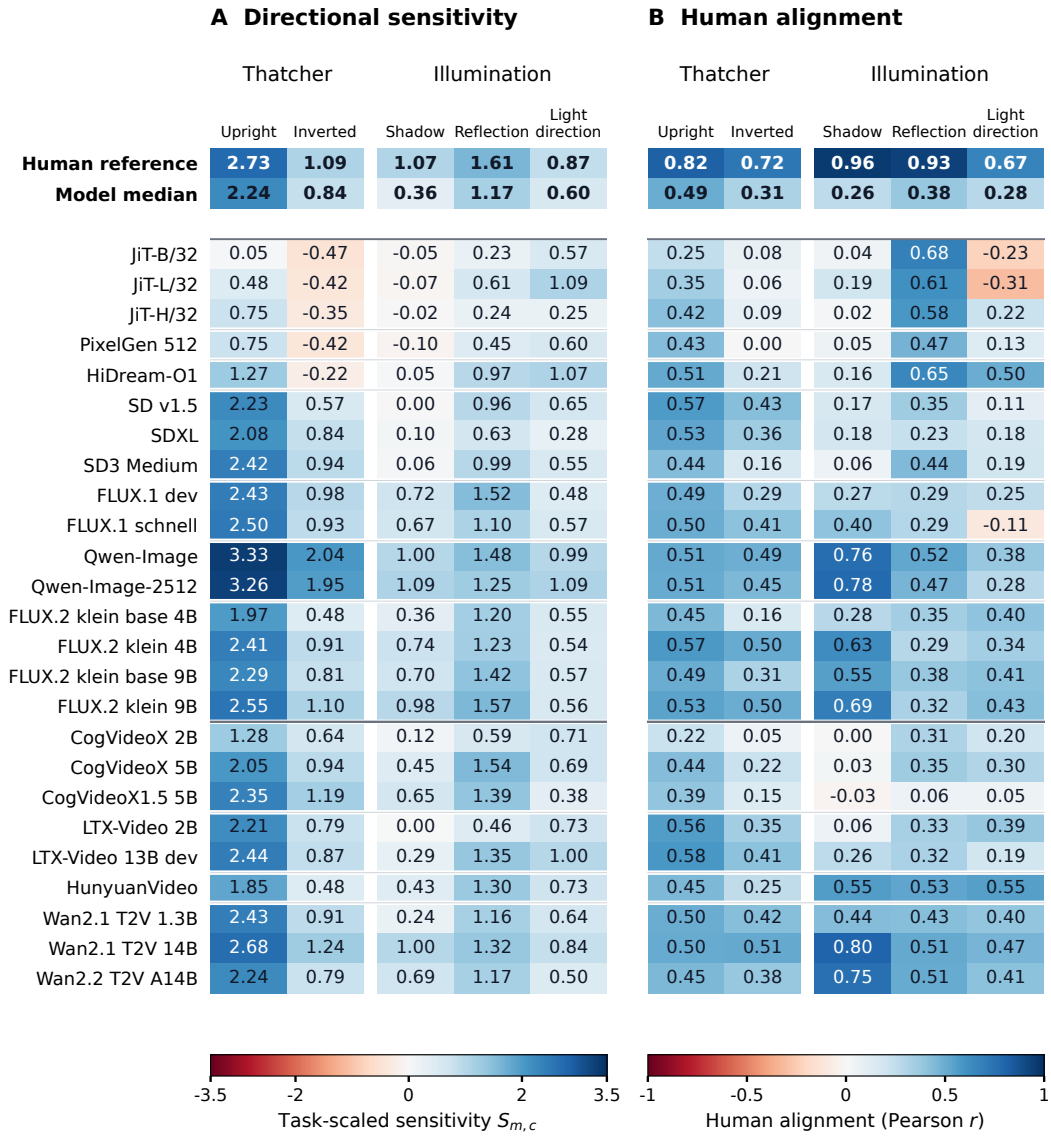

Figure S12: **Model-wise sensitivity and human alignment across conditions.** (A) Directional sensitivity, $S_{m,c} = \bar{u}_{m,c}/s_{m,\text{task}}$: the condition mean loss difference divided by the model's sample SD over the full task. (B) Human alignment: Pearson correlation between model loss differences and human paired rating differences within each condition. The top rows show human references and the median across all 25 models; human alignment is the mean raw split-half correlation across 10,000 disjoint participant splits.

H SENSITIVITY AND HUMAN ALIGNMENT ACROSS INDIVIDUAL CONDITIONS

Figure S13 shows sensitivity–alignment associations across the 25 models for each stimulus condition. Pointwise 95% intervals combine family, participant, and scene resampling (Appendix E).

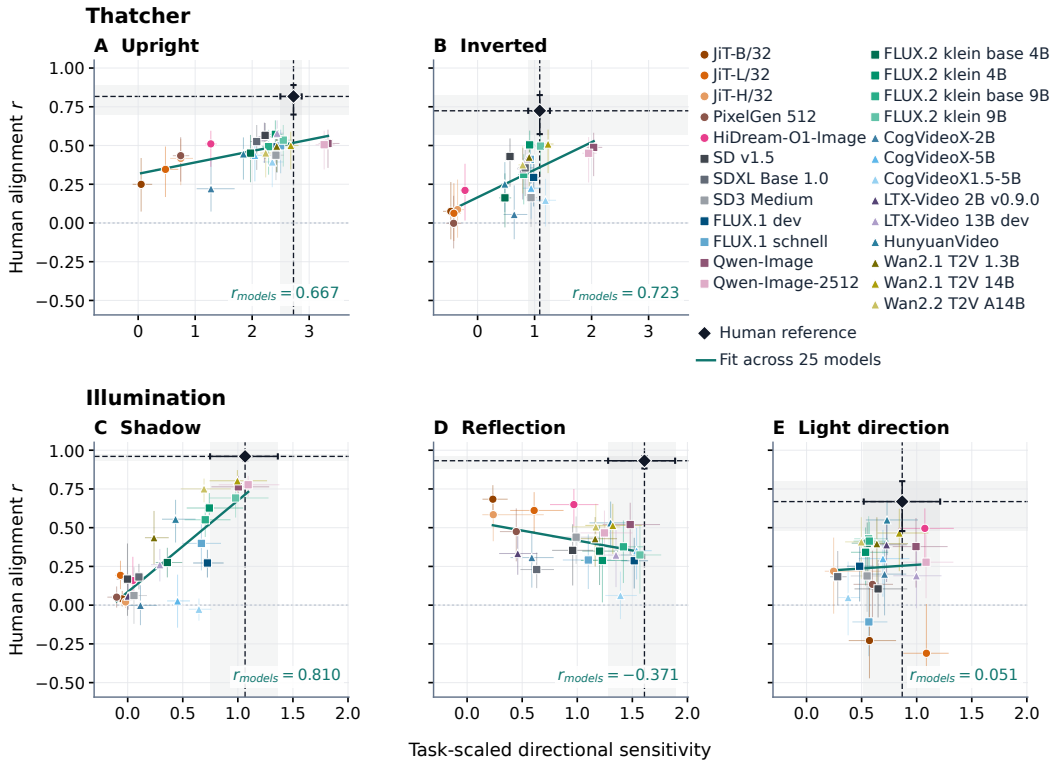

Figure S13: **Sensitivity and human alignment across stimulus conditions.** Directional sensitivity versus human alignment (r) for 25 generative models across individual conditions: upright Thatcher (A), inverted Thatcher (B), cast shadow (C), surface reflection (D), and light direction (E). Error bars and shaded bands indicate 95% bootstrap intervals (Appendix E). Black diamonds and dashed crosshairs mark human sensitivity and mean split-half reliability. Green lines show linear fits across models.

Cross-model sensitivity–alignment associations varied across stimulus conditions (Figure S13). Positive associations were robust for inverted faces ($r = .723$, 95% CI $[.118, .923]$) and cast shadows ($r = .810$, $[.424, .960]$), where models with stronger directional penalties also exhibited closer alignment with continuous human ratings. In contrast, this relationship became weak or inconsistent in the remaining conditions: while upright faces showed a positive point estimate ($r = .667$, $[-.215, .904]$), its confidence interval was wide, and associations were absent or negative for light direction ($r = .051$, $[-.417, .570]$) and surface reflections ($r = -.371$, $[-.704, .364]$).

I Condition-specific evaluation against non-generative baselines

We evaluate baseline comparisons separately within each stimulus condition. To facilitate direct comparison without confusing model shifts across conditions, we retain the models selected over each entire task: the top five generative models, top five encoders, and all IQA metrics. Figure S14 displays the resulting marginal alignment and bidirectional partial correlations.

Across individual conditions, generative models consistently retained positive residual correlations with human judgments ($r(h, G \mid X)$) after controlling for each non-generative baseline (Figure S14, bottom panels). For upright faces, cast shadows, and light direction, generative scores accounted for substantial human-aligned variance beyond visual encoders and standard quality metrics, whereas baseline residual associations were comparatively low. Under face inversion, residual correlations were attenuated for both generative models and baselines, consistent with the overall reduction in perceptual sensitivity. For surface reflections, both generative models and visual encoders retained

positive residual associations ($r(h, G \mid \text{Encoder}) \approx .346$, $r(h, \text{Encoder} \mid G) \approx .584$), indicating that both representations capture distinct aspects of reflection naturalness.

Because condition-level aggregates combine multiple scene properties, they do not isolate which factors drive these agreements and discrepancies. In Appendix J, we break down the Illumination stimuli into specific scene factors (including object shape, orientation, and surface roughness) to examine differences across model classes.

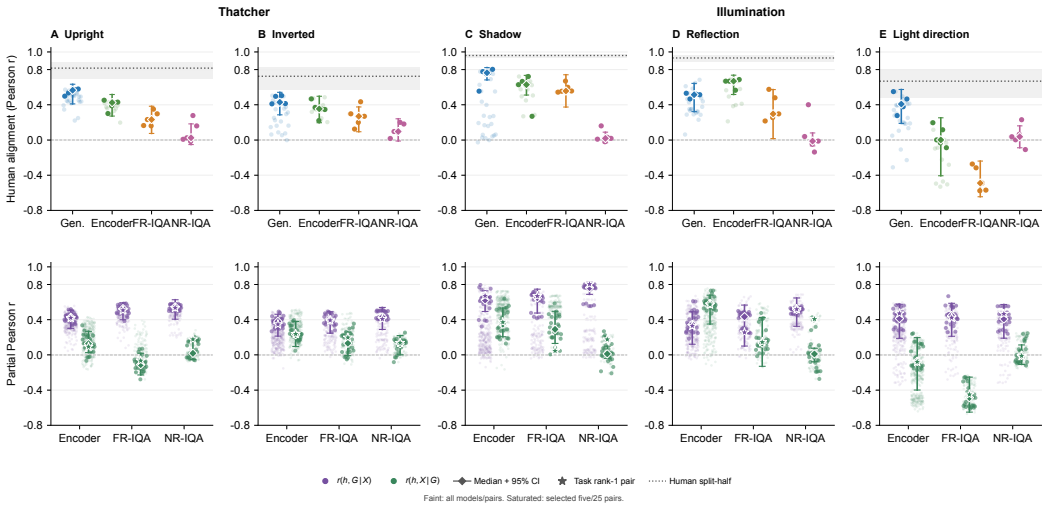

Figure S14: **Condition-level alignment and partial correlation with baseline models.** Panels show Thatcher (**A**, upright; **B**, inverted) and Illumination (**C**, shadow; **D**, reflection; **E**, light direction), evaluated with the top-five models and encoder layers selected over each full task. Top panels show marginal human alignment (faint points: all candidates; saturated points: selected models; diamonds and error bars: selected median and 95% bootstrap CI; dotted lines and bands: mean human split-half reliability and its 95% interval). Bottom panels show bidirectional partial correlations (faint points: all pairs; saturated points: selected 5×5 pairs; stars: rank-1 pair; diamonds and bars: selected median and 95% CI). Purple circles show $r(h, G \mid X)$ and green circles show $r(h, X \mid G)$. Bootstrap intervals account for task-level layer and model reselection (Appendix E.3).

J FACTOR-LEVEL STRUCTURE OF ILLUMINATION JUDGMENTS

While overall evaluations measure aggregate alignment across the full stimulus set, they obscure which specific scene properties accompany human–model consistency or divergence. We therefore examined how perceived unnaturalness and model predictions vary across the object shapes, rotations, colors, and physical parameters of the Illumination stimuli (Appendix A). We evaluate the task-pooled top-five models (Section 3.4): generative models Wan2.1 T2V 14B, Qwen-Image, Wan2.2 T2V A14B, Qwen-Image-2512, and HunyuanVideo (Table S1); and encoders PE-Core B/16, PE-Core G/14, CLIP ViT-B/16, SigLIP 2 Base/16, and Jina CLIP v2 (Table S3). All five metrics from each IQA battery are also retained. The same model sets are used in the condition-level comparison in Figure S14.

J.1 FACTOR PROFILES AND COMPARISON METRIC

For each of the 13 scene factors (Appendix A), a profile plots the mean paired response at each factor level, marginalizing over other scene parameters. Each human or model score is divided by its own sample standard deviation across all 288 Illumination pairs ($\mathrm{ddof} = 1$), without subtracting its mean; factor-level means therefore retain the original zero point ($S = \bar{u}/s_{\mathrm{task}}$, Section 3.3). FR-IQA metrics are converted to distance forms before task-SD scaling: PSNR becomes unit-range pixel MSE ($10^{-\mathrm{PSNR}/10}$), SSIM and VIF become $1 - \mathrm{SSIM}$ and $1 - \mathrm{VIF}$, and LPIPS and DISTS retain their native distance values. To quantify human–model agreement, we compute the mean squared error (MSE) between model and human profiles:

$$\mathrm{MSE}_{mf} = \frac{1}{K_f} \sum_{\ell=1}^{K_f} (y_{mf\ell} - y_{hf\ell})^2, \qquad \mathrm{MSE}_m = \sum_f v_f \mathrm{MSE}_{mf}, \tag{7}$$

where $y_{mf\ell}$ and $y_{hf\ell}$ are the scaled model and human means at level ℓ, K_f is the number of levels in factor f, and weights v_f weight each of the three operations equally ($1/3$ per operation, divided evenly among its factors). Figures S16–S18 display the level-wise profiles, and Figure S15 summarizes the resulting MSE scores. Human profile means include 95% participant bootstrap intervals (Appendix E), while category summaries show medians across the five models in each class, displaying model spread rather than confidence intervals.

J.2 HUMAN VARIATION ACROSS SCENE FACTORS

Geometry and rotation. Human unnaturalness judgments varied markedly across specific shapes and transformations. In Light direction, mirroring the central sphere elicited strong perceived unnaturalness (1.805), whereas the same manipulation of the complex torus knot was virtually unnoticed (0.014; Figure S18, L1). Similarly, rotating the torus knot produced negligible unnaturalness in both Shadow (-0.077 and 0.013) and Reflection (0.018 and 0.089). By contrast, rotations of familiar objects with distinct axes were highly salient: Shadow scores reached 0.714 for the torus ($0° \leftrightarrow 45°$) and 1.462 for the teapot ($90° \leftrightarrow 180°$), while Spot cow rotations in Reflection scored 1.145 and 1.540 (Figures S16 and S17, S2/R2). These patterns indicate that within these scenes, human sensitivity to object–effect inconsistencies is strongly modulated by object geometry.

Shape swaps, in contrast, produced consistently salient unnaturalness across all tested pairs, spanning 1.261–1.913 in Shadow and 1.784–2.155 in Reflection (S1/R1). Averaged across physical settings, variation among different object pairs was substantially narrower than the spread observed across rotation conditions.

Shadow softness, floor roughness, and height. Physical scene parameters also produced systematic and intuitive effects. Shadow inconsistencies were far more salient with hard than soft cast shadows (1.731 vs. 0.400; S4). Reflection anomalies were similarly more conspicuous on smooth than rough surfaces (2.166 vs. 1.054; R5), and reflection color shifts elicited consistently high unnaturalness across both objects even for subtle hue differences (1.730–2.416; R3). Placing objects closer to the ground increased perceived anomaly in Shadow (1.224 vs. 0.908) and Reflection (1.773 vs. 1.447). In Light direction, mean perceived unnaturalness was similar in the low-height scenes with cast shadows and the high-height scenes without them (0.850 vs. 0.887; L3). Finally, for Light direction, diffuse materials elicited higher unnaturalness than metallic surfaces (1.063 vs. 0.675),

and lighting from a more laterally tilted angle ($30°$) elicited greater unnaturalness than from a more overhead angle ($15°$; 1.001 vs. 0.736).

J.3 Where the model profiles agree and disagree

Generative errors were smaller for several geometric profiles. In rotation, generative models broadly tracked human sensitivity (median $\mathrm{MSE} = 0.061$ in Shadow, 0.324 in Reflection), whereas encoders (0.418 and 0.465), FR-IQA (1.041 and 0.645), and NR-IQA (0.589 and 0.877) showed larger discrepancies. As shown in the level-wise plots (Figures S16–S17), generative models broadly reproduced the distinction between subtle torus-knot rotations and conspicuous teapot or Spot rotations. Similarly, in Light direction (L1), generative models captured the human contrast between strong sensitivity on spheres and insensitivity on the torus knot (generative median $\mathrm{MSE} = 0.170$ vs. encoder 0.632, FR-IQA 1.058, and NR-IQA 1.193). Encoders and FR-IQA metrics often responded strongly to the light-direction manipulation of the torus knot despite human judgments being near baseline.

Encoders had lower median error than generative models for reflection shapes (R1) and color shifts (R3). Whereas generative models had lower error for shadow shape swaps (S1: $\mathrm{MSE} = 0.419$ vs. encoder 1.540), this ordering reversed in Reflection (R1): humans perceived similarly strong unnaturalness across object pairs (1.784–2.155), whereas generative scores varied more across object geometries, incurring higher error ($\mathrm{MSE} = 0.753$) than FR-IQA (0.403) and encoders (0.456). For reflection color shifts (R3), observers reported marked unnaturalness even for subtle hue differences (1.730 and 2.052), but median generative responses were much smaller (0.480 and 0.431, respectively) and rose more strongly for large shifts. This produced the largest generative profile error (median $\mathrm{MSE} = 1.292$), compared with 1.037 for encoders. These factor patterns identify specific mismatches but are not a decomposition of the condition-level partial correlations, where encoders retained a substantial residual association for surface reflections (Figure S14).

For shadow softness (S4), generative models closely tracked human sensitivity (median $\mathrm{MSE} = 0.043$), whereas FR-IQA metrics (0.161) and visual encoders (0.922) exhibited larger discrepancies. For floor roughness in Reflection (R5), generative and encoder errors were similar (both 0.171), as was the FR-IQA error (0.181). Finally, for object height in Light direction (L3), where neither humans (0.850 vs. 0.887) nor models showed substantial variation, errors remained relatively low for both FR-IQA metrics ($\mathrm{MSE} = 0.016$) and generative models (0.104).

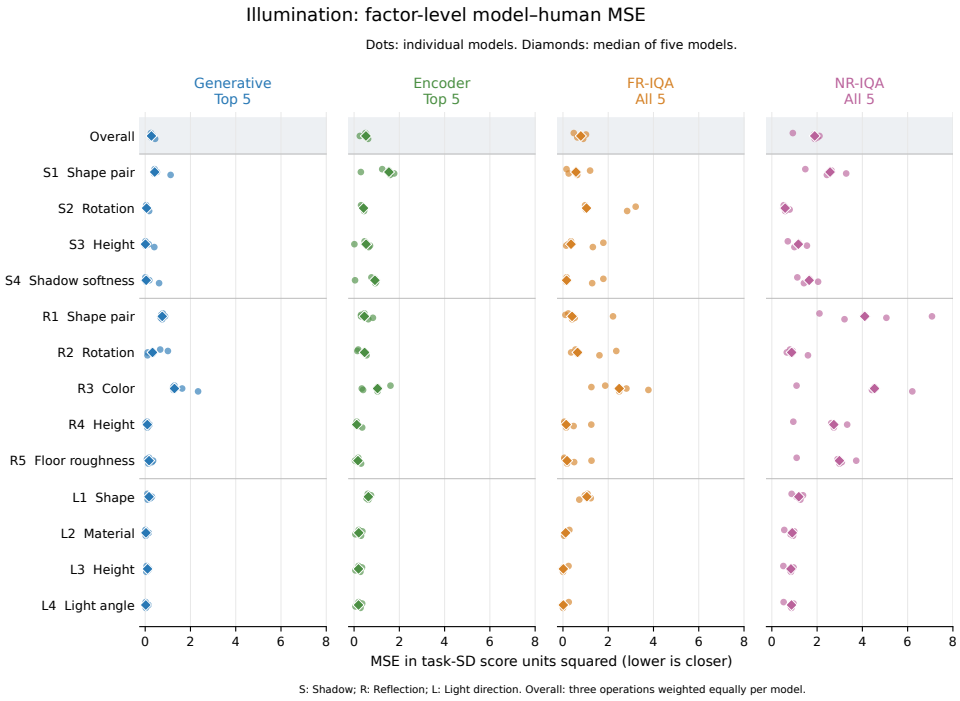

Figure S15: **Factor-profile errors across model categories.** Columns separate categories and rows separate factors. Dots show individual models (five per category: generative models, visual encoders, FR-IQA, and NR-IQA); diamonds indicate category medians. Models and encoder layers are fixed by task-pooled alignment. The spread describes model variation, not confidence intervals. Overall displays the aggregate MSE across all 13 factor profiles (MSE_m, Eq. 7) for each model. Factor codes correspond to the detailed profiles in Figures S16–S18.

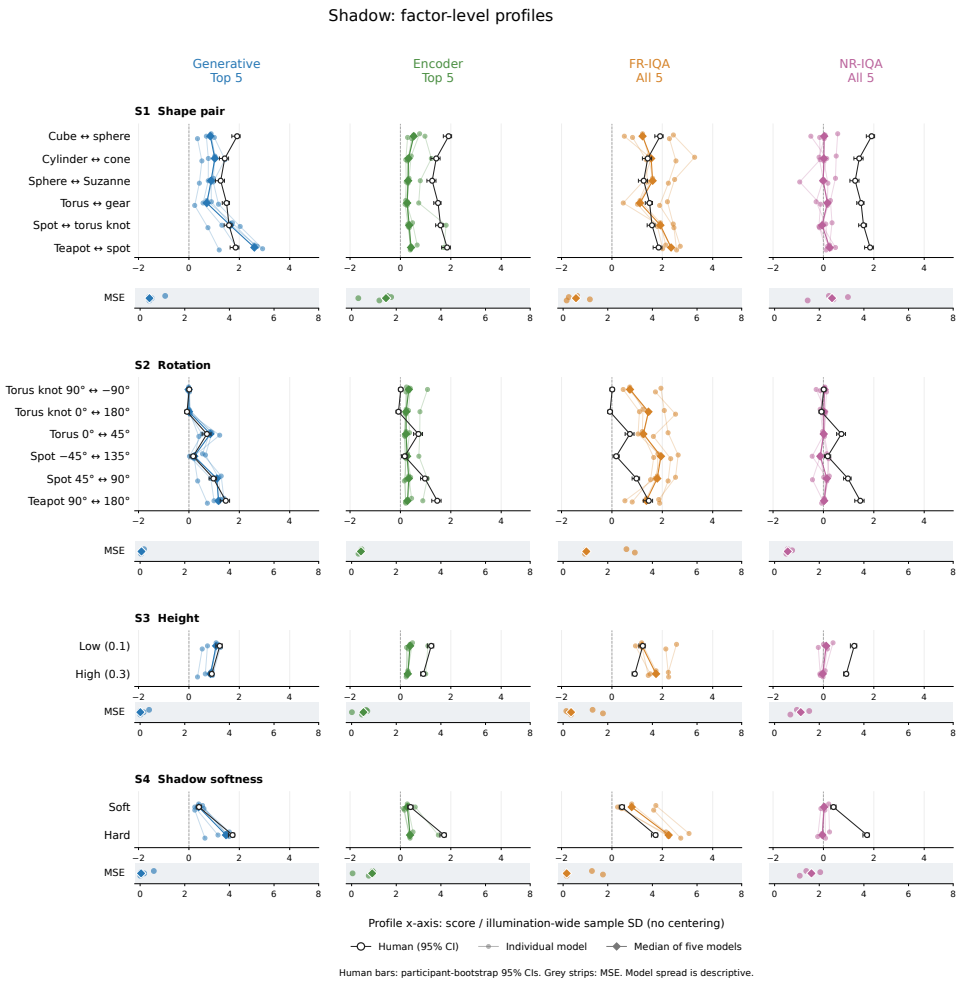

Figure S16: **Shadow: factor profiles and model error.** Columns separate the four model categories (generative models, visual encoders, FR-IQA, and NR-IQA). Black open circles and horizontal bars show human means and 95% bootstrap intervals (Appendix E), repeated across columns. Colored dots show individual models, faint lines connect the same model across factor levels, and filled diamonds denote category medians. Gray strips show corresponding factor MSEs. Responses are scaled by task standard deviation without centering (encoder distances are unsigned), with model sets and encoder layers held constant across factors.

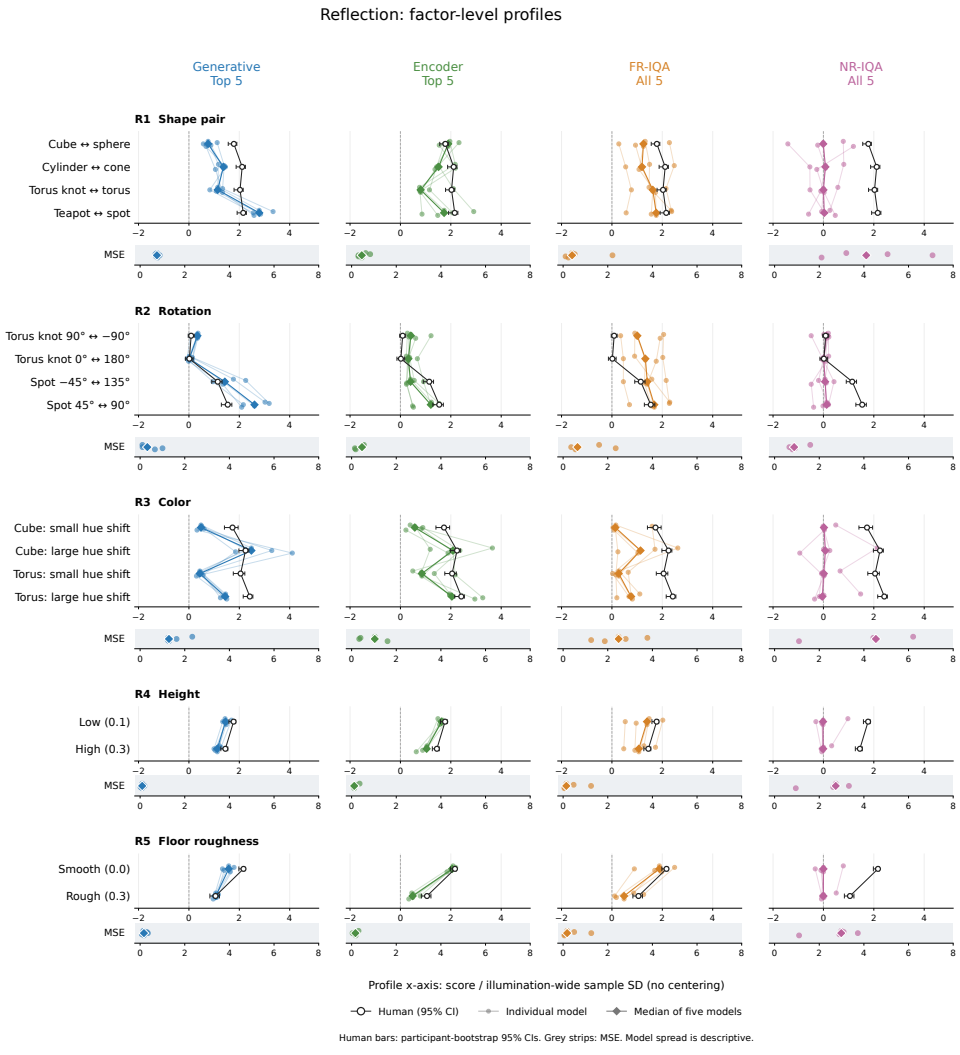

Figure S17: **Reflection: factor profiles and model error.** Plot conventions match Figure S16.

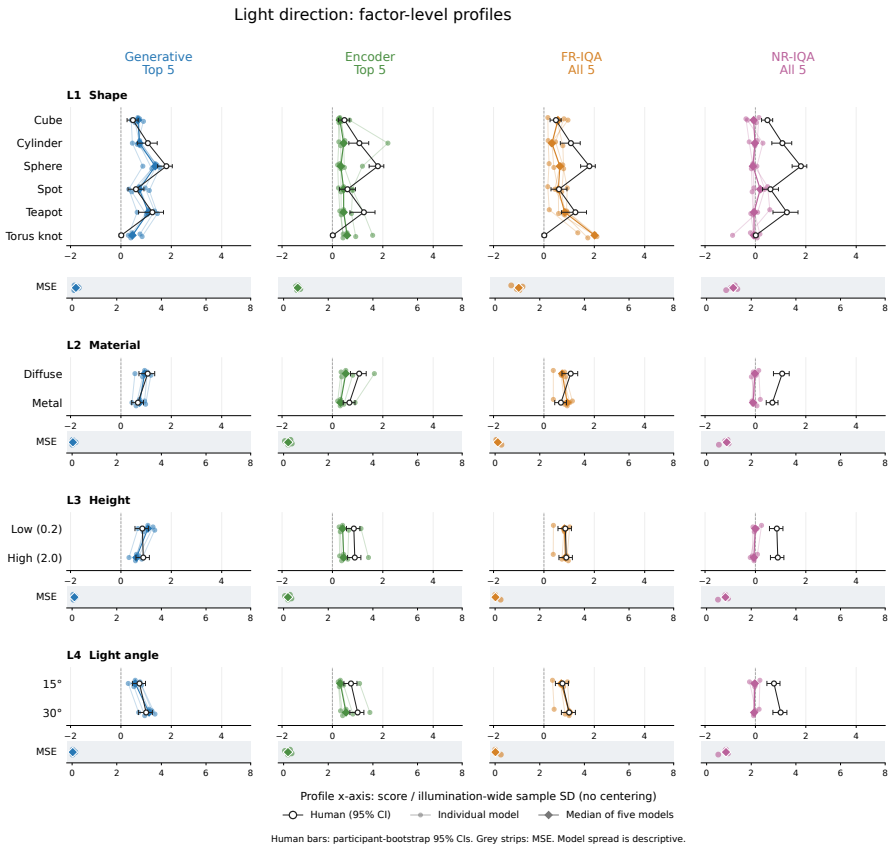

Figure S18: **Light direction: factor profiles and model error.** Plot conventions match Figure S16.

K EFFECTS OF MODEL CAPACITY ON SENSITIVITY AND HUMAN ALIGNMENT

We evaluated how directional sensitivity and human alignment relate to model capacity across the 25 generative models (Figure S19). For generative models, model size is defined as the parameter count of the denoising or flow backbone network used to compute native scores (ranging from 0.14B to 20B parameters).

Larger generators generally tended to show both stronger directional sensitivity and closer human alignment, though these associations varied across tasks and metrics. Log parameter count correlated positively with directional sensitivity (Thatcher: Pearson $r = .741$, Spearman $\rho = .685$; Illumination: $r = .831$, $\rho = .841$) and human alignment (Thatcher: $r = .627$, $\rho = .375$; Illumination: $r = .803$, $\rho = .796$). The weaker rank correlation for Thatcher alignment ($\rho = .375$) suggests that while larger models were generally more sensitive, human alignment did not follow model size equally closely across all evaluated models.

To provide an empirical reference over an overlapping parameter range, we also evaluated all 12 frozen vision and multimodal encoders at the layer selected for each task (Table S2; 0.02B to 8.1B parameters; Figure S19, open green diamonds). Encoder size denotes the full scoring network, not the prefix ending at the selected layer. The same standardization (mean feature distance divided by its standard deviation across the task) was applied to unsigned distances, which measure representational divergence rather than directional unnaturalness. At the selected layers, encoder sensitivity was moderately negatively associated with log parameter count on Thatcher and weakly associated on Illumination (Thatcher: $r = -.308$, $\rho = -.399$; Illumination: $r = .090$, $\rho = .028$). Encoder human alignment was positively associated with size on Thatcher ($r = .380$, $\rho = .322$), but weakly associated on Illumination ($r = -.110$, $\rho = -.098$). Thus, whereas larger generators show both stronger directional sensitivity and closer human alignment, scaling encoder size does not automatically improve either metric, particularly on Illumination.

Condition-specific scaling trajectories. Decomposing this relationship by stimulus condition (Figures S20 and S21) reveals that model capacity affects directional sensitivity and human alignment through distinct trajectories. Directional sensitivity scaled positively with parameter count across nearly all conditions (Thatcher upright: Pearson $r = .760, \rho = .694$; inverted: $r = .697, \rho = .641$; cast shadow: $r = .790, \rho = .846$; surface reflection: $r = .783, \rho = .729$; light direction: $r = .244, \rho = .191$), indicating that larger backbones systematically acquire stronger penalties for relational alterations across diverse visual domains. In contrast, human alignment exhibited pronounced domain selectivity: while alignment scaled with capacity for cast shadows ($r = .649, \rho = .688$) and light direction ($r = .610, \rho = .526$), it showed no positive association with model size for surface reflections ($r = -.256, \rho = -.104$) and saturated early for upright faces ($r = .516, \rho = .378$).

Crucially, the absence of alignment scaling for surface reflections cannot be attributed to human rating noise, as participants exhibited high split-half reliability in this condition ($r = .932$; Appendix E.2). Instead, it reflects a genuine divergence between the statistical features captured by generative loss minimization (which readily penalizes surface modifications) and the physical or perceptual criteria humans employ to judge reflection naturalness.

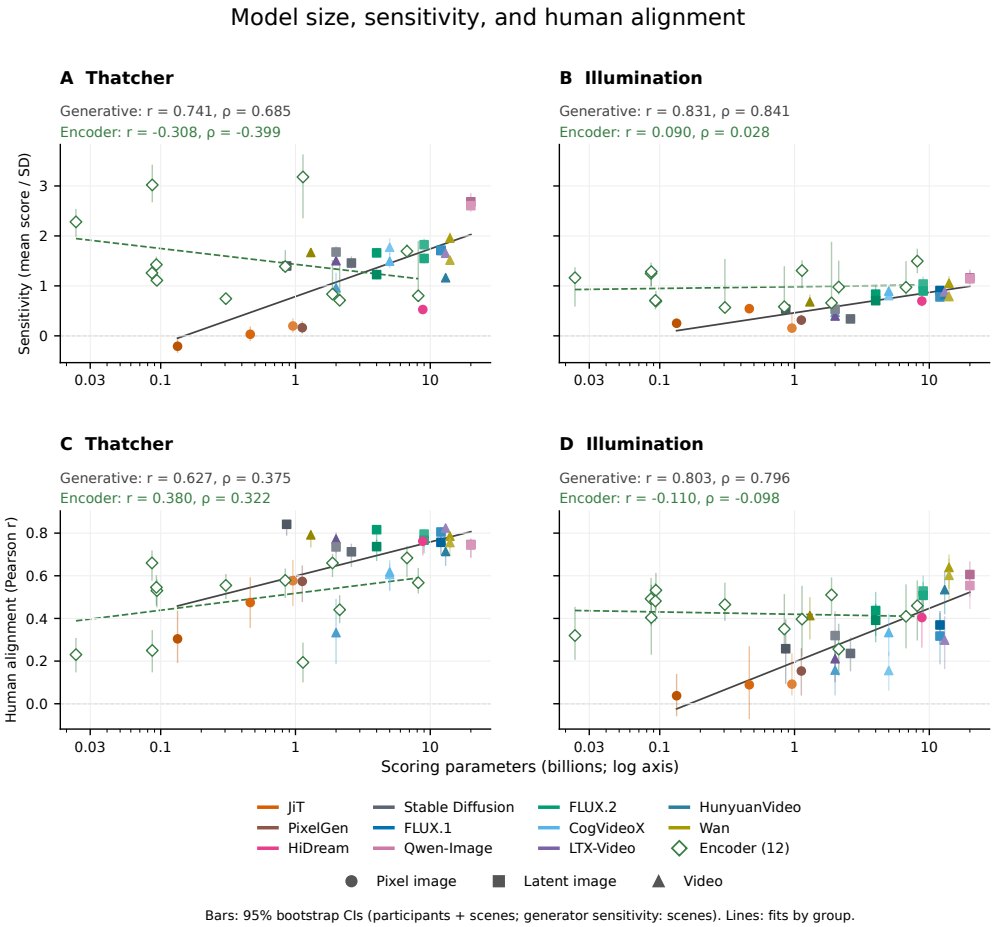

Figure S19: **Model size, sensitivity, and human alignment.** (**A, B**) Directional sensitivity and (**C, D**) human alignment (r) versus scoring parameter count for 25 generators and 12 encoders. Solid and dashed lines show linear fits against log parameter count for generators and encoders, respectively; annotations report Pearson r and Spearman ρ. Error bars indicate 95% bootstrap intervals (Appendix E). Colors and markers match Figure 3 and the bottom legend.

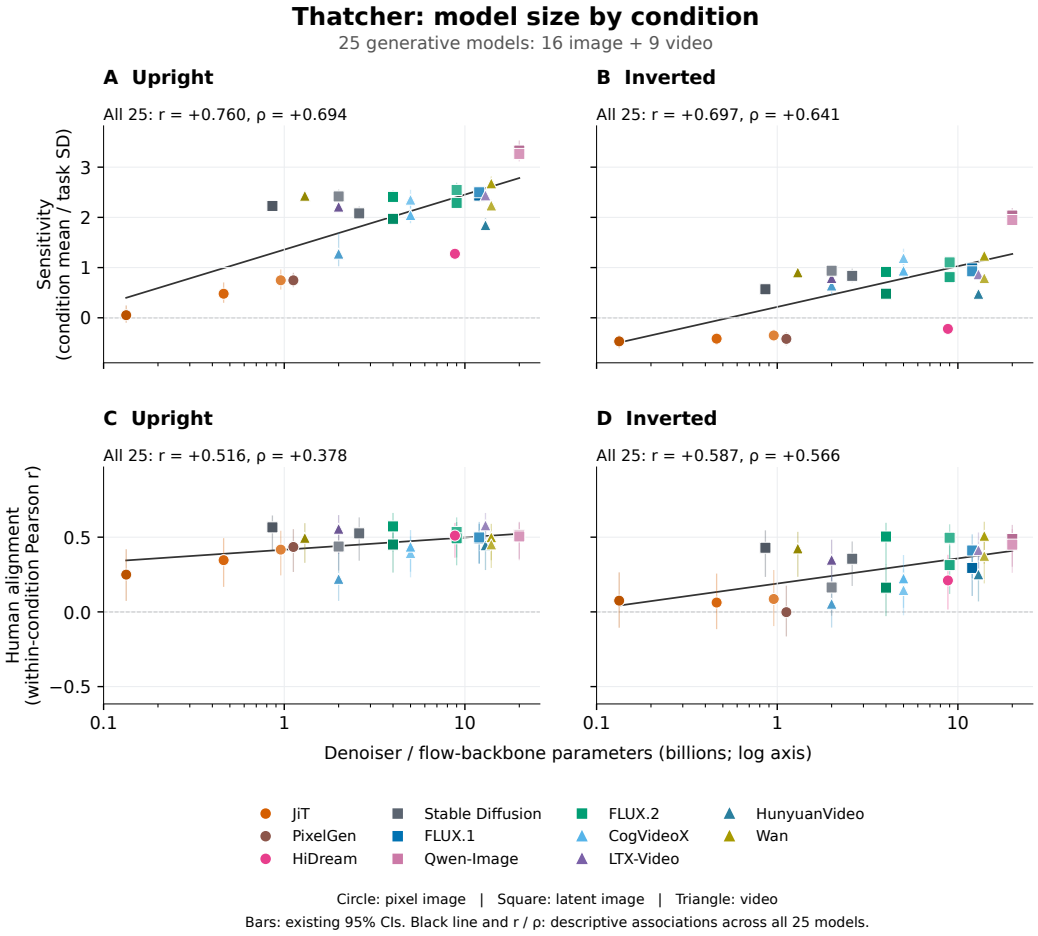

Figure S20: **Model size and condition-specific metrics in the Thatcher task.** (**A, B**) Directional sensitivity and (**C, D**) human alignment (r) versus scoring parameter count for upright and inverted faces across 25 generators. Solid lines show linear fits against log parameter count; annotations report Pearson r and Spearman ρ. Error bars indicate 95% bootstrap intervals (Appendix E). Colors and markers match Figure 3 and the bottom legend.

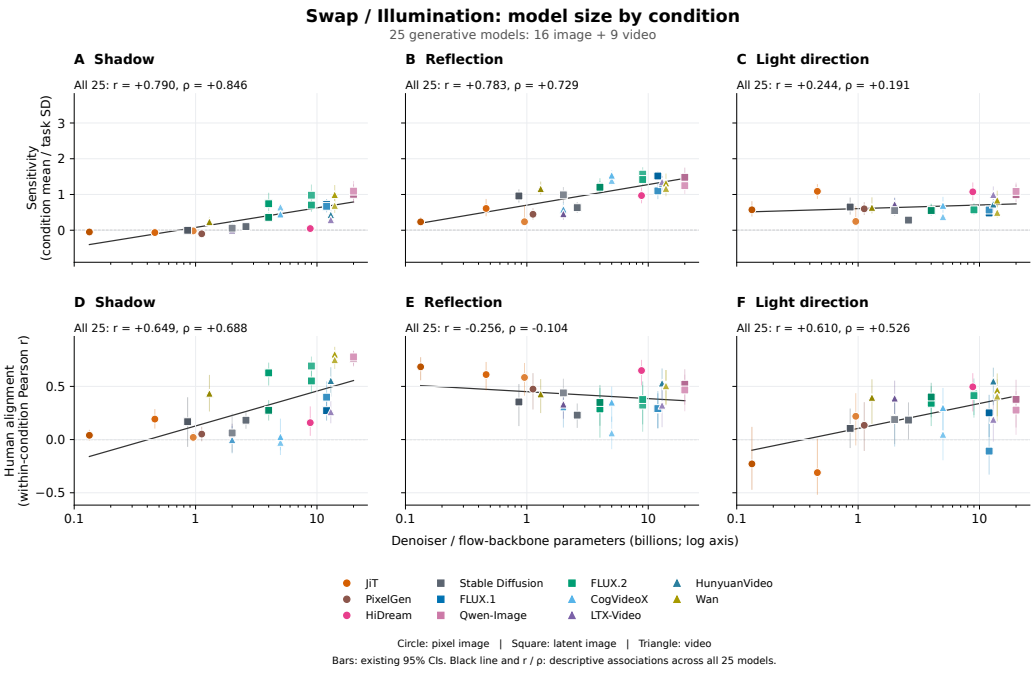

Figure S21: **Model size and condition-specific metrics in the Illumination task.** (**A–C**) Directional sensitivity and (**D–F**) human alignment (r) versus scoring parameter count for cast shadow, surface reflection, and light direction across 25 generators. Solid lines show linear fits against log parameter count; annotations report Pearson r and Spearman ρ. Error bars indicate 95% bootstrap intervals (Appendix E). Colors and markers match Figure S20 and the bottom legend.

L PUBLISHED GENERATION EVALUATIONS AND HUMAN ALIGNMENT

We conducted a post-hoc comparison with published generation evaluations to ask whether models' human alignment in our tasks is associated with broader generation performance, beyond what can be explained by overall directional sensitivity.

L.1 EVALUATION CHOICE AND MODEL MATCHING

We evaluated a subset of 11 image and seven video models from our primary sample for which corresponding public benchmark records were available (Table S4). Image models are evaluated on Artificial Analysis text-to-image Elo (Artificial Analysis, 2026), derived from blinded pairwise human judgments across shared prompts. Video models are evaluated on the automated VBench Quality score (Huang et al., 2024), which assesses visual and temporal fidelity using pretrained vision models and motion algorithms (spanning consistency, flickering, smoothness, aesthetics, and imaging quality).[1] Model correspondence was established from published model names and official checkpoint links; models without corresponding public evaluation records (e.g., JiT, PixelGen, LTX) were excluded.

Table S4: **Models matched to public generation evaluations.** Parameters refer to the scoring backbone in billions. The same IDs, colors, and markers identify models in Figures S22 and S23. Published names do not guarantee identical checkpoint revisions or generation pipelines.

ID	Evaluated model	B	Published entry	Score
Image models: Artificial Analysis human-preference Elo				
P05	HiDream-O1-Image	8.805	HiDream-O1-Image	979
L01	Stable Diffusion v1.5	0.86	Stable Diffusion 1.5	468
L02	SDXL Base 1.0	2.6	Stable Diffusion XL 1.0	688
L03	Stable Diffusion 3 Medium	2	Stable Diffusion 3 Medium	726
L04	FLUX.1 dev	12	FLUX.1 [dev]	842
L05	FLUX.1 schnell	12	FLUX.1 [schnell]	807
L06	Qwen-Image	20	Qwen Image	883
L08	FLUX.2 klein base 4B	4	FLUX.2 [klein] Base 4B	776
L09	FLUX.2 klein 4B	4	FLUX.2 [klein] 4B	864
L10	FLUX.2 klein base 9B	9	FLUX.2 [klein] Base 9B	902
L11	FLUX.2 klein 9B	9	FLUX.2 [klein] 9B	947
Video models: VBench Quality (%)				
V01	CogVideoX-2B	2	CogVideoX-2B (Diffusers)	82.480
V02	CogVideoX-5B	5	CogVideoX-5B (Diffusers)	83.050
V03	CogVideoX1.5-5B	5	CogVideoX1.5-5B	82.720
V06	HunyuanVideo	13	HunyuanVideo (Open-Source Version)	85.090
V07	Wan2.1 T2V 1.3B	1.3	Wan2.1-T2V-1.3B (2025-05-03)	85.230
V08	Wan2.1 T2V 14B	14	Wan2.1-T2V-14B	85.590
V09	Wan2.2 T2V A14B	14	Wan2.2-T2V-A14B (w/o prompt-optimized)	85.030

L.2 ASSOCIATIONS BEFORE AND AFTER SENSITIVITY ADJUSTMENT

We retained the same 11 image and seven video models for both tasks. Let Q denote the public generation score, A the model's human alignment ($A_m = r_m$), and S directional sensitivity (S_m). Before adjustment, the model-level Spearman correlations between Q and A were .191 and .827 for image models on Thatcher and Illumination, respectively, and .929 and .821 for video models (Figure S22; Table S5).

[1] Snapshots retrieved September 11, 2026 from https://artificialanalysis.ai/image/leaderboard/text-to-image and https://huggingface.co/spaces/Vchitect/VBench_Leaderboard.

Models (n)	Task	$\rho(Q, A)$	$\rho(Q, A \mid S)$	$\rho(Q, S \mid A)$
Image (11)	Thatcher	.191	.182	.201
	Illumination	.827	.670	−.045
Video (7)	Thatcher	.929	.900	.122
	Illumination	.821	.778	.218

Table S5: **Generation evaluations and human alignment.** Spearman associations of generation evaluations Q, alignment A, and sensitivity S; variables after | are controlled. Scatter plots are in Figures S22–S24.

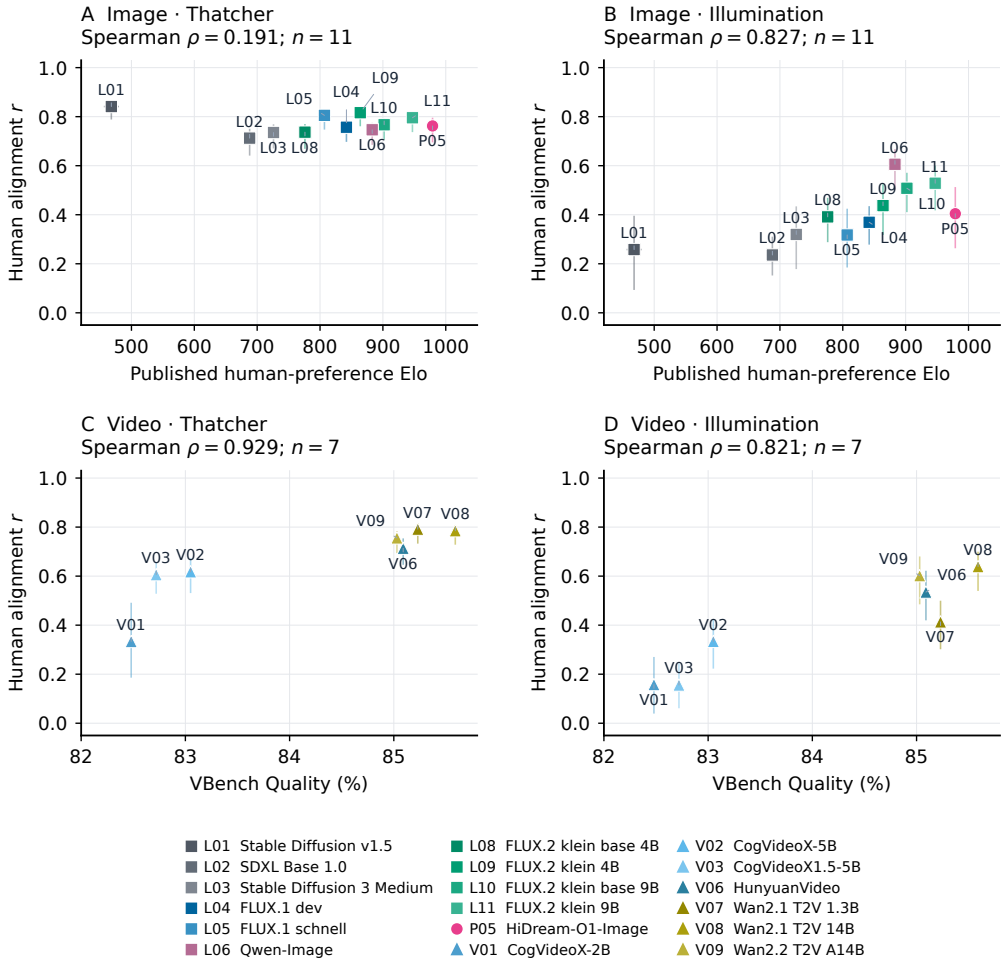

Figure S22: **Published generation evaluations and human alignment before adjustment.** Image models (A, B; $n = 11$) and video models (C, D; $n = 7$), with Thatcher on the left and Illumination on the right. Each point represents one model; IDs, colors, and markers match Figure S23. The vertical axis is human alignment (r). Vertical bars show 95% participant-and-scene bootstrap CIs for each model; horizontal bars in A and B show published Elo intervals.

To control for sensitivity, we transformed Q, A, and S to average ranks within each image/video task panel, separately regressed the ranks of Q and A on an intercept and the rank of S, and correlated the two residuals using Pearson correlation, without reranking them. The resulting partial Spearman coefficients were .182 and .670 for image models and .900 and .778 for video models (Figure S23). Thus, positive associations remained on Illumination for image models and both tasks for video models, whereas the association on Thatcher for image models was weak.

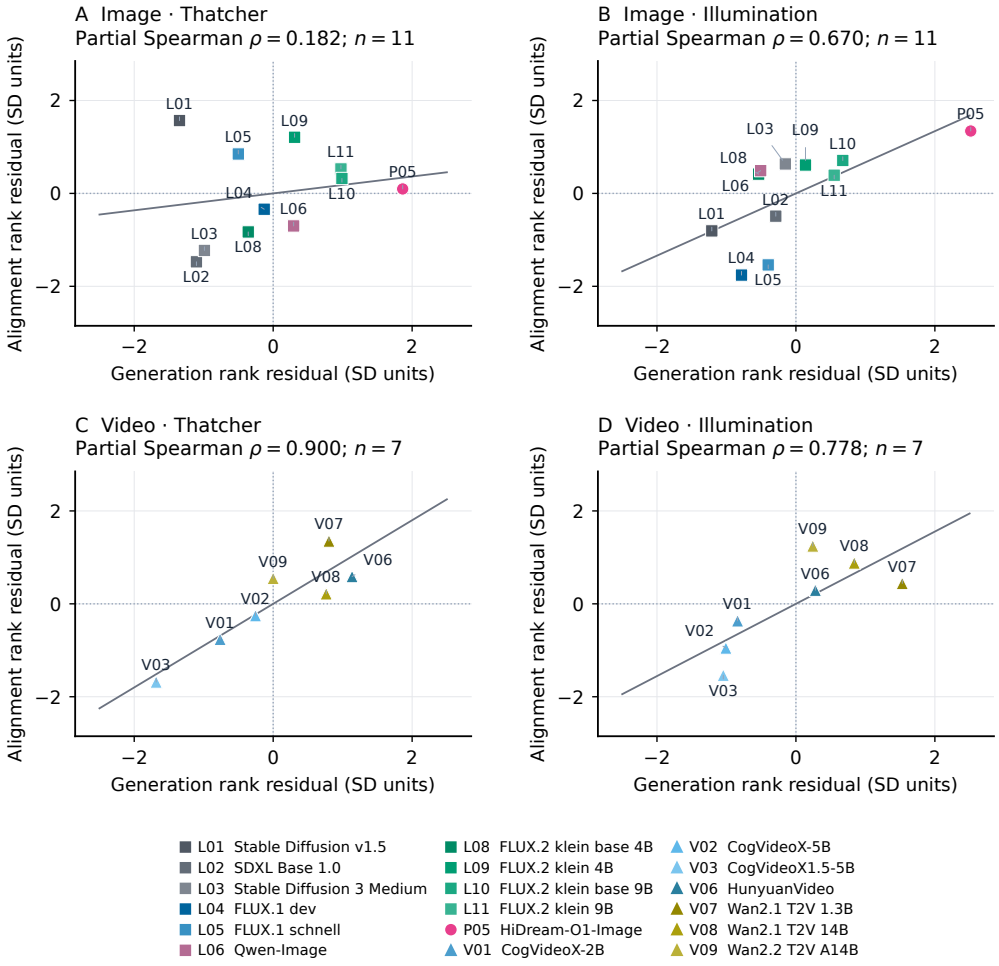

Figure S23: **Generation evaluations and human alignment after controlling for directional sensitivity.** The same models and panel arrangement as Figure S22. Both axes show residuals from regressing the corresponding model ranks on sensitivity rank, divided by their own sample standard deviations. Zero denotes agreement with the sensitivity-based regression prediction. Gray lines show least-squares fits, whose slope equals the annotated partial Spearman correlation. These are the sensitivity-adjusted values in Table S5.

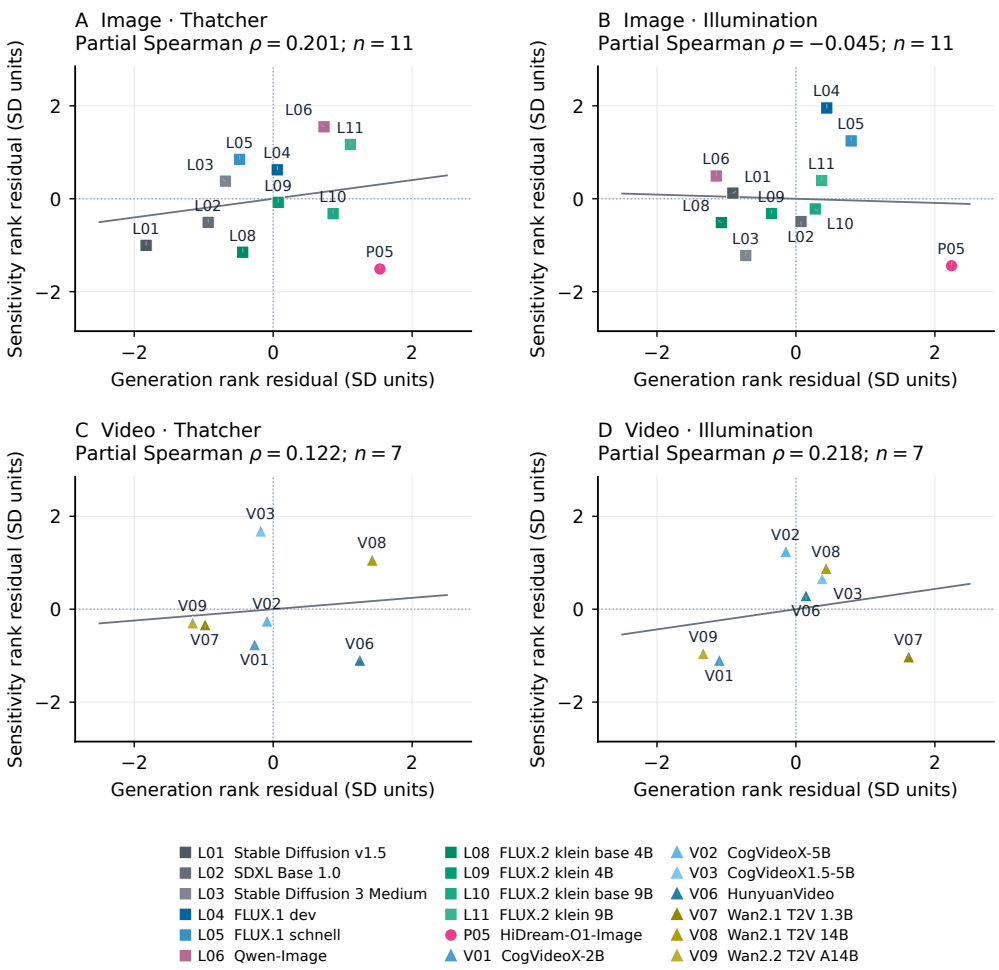

Figure S24: **Generation evaluations and sensitivity after controlling for human alignment.** Same format, models, and conventions as Figure S23, showing standardized rank residuals of generation evaluations (horizontal) and directional sensitivity (vertical) after regressing out human alignment rank. These are the alignment-adjusted values in Table S5.

L.3 Sensitivity after accounting for human alignment

We also examined the reciprocal association by separately regressing the ranks of Q and S on an intercept and the rank of A, using the same matched models and residual-correlation procedure. Partial Spearman correlations were .201 and $-.045$ for image models on Thatcher and Illumination, respectively, and .122 and .218 for video models (Table S5; Figure S24). These weak residual associations reinforce the interpretation that generation performance tracks fine-grained human alignment rather than broad violation sensitivity alone, although the modest sample sizes warrant cautious interpretation.

L.4 Additional size adjustment and interpretation

As a supplementary check, we added scoring-backbone parameter count P (Table S4) to the sensitivity adjustment. Partial Spearman correlations controlled for the ranks of both S and P. Positive associations remained on Illumination for image models and both tasks for video models, while the image-Thatcher association remained weak (Table S6). Controlling for parameter count did not attenuate these associations, suggesting that the correspondence between perceptual alignment and practical generation quality persists beyond simple scaling effects.

Table S6: **Generation-evaluation associations with sensitivity and size adjustment.** All coefficients relate generation evaluation Q to human alignment A. S is directional sensitivity and P is scoring-backbone parameter count. Partial Spearman coefficients control ranks of the indicated covariates. Entries are descriptive point estimates in the same matched sample as Table S5.

			Partial Spearman	
Models	Task	n	S	S, P
Image	Thatcher	11	.182	.222
	Illumination	11	.670	.783
Video	Thatcher	7	.900	.913
	Illumination	7	.778	.914

M Spatial loss-difference maps across generative models

Figures S25–S27 display condition-averaged loss-difference maps across all 25 generative models (16 image and 9 video models). At each native spatial position, we compute the paired difference by subtracting original-image loss from modified-image loss, averaged uniformly across timesteps and noise draws (and latent frames for video models). For model m, task d, and condition c, the mean loss-change map is defined as:

$$\overline{M}_{m,d,c}(x,y) = \frac{1}{N_{d,c}\, s_{m,\text{task}}} \sum_{i \in I_{d,c}} \left[L^{\text{mod}}_{m,i}(x,y) - L^{\text{orig}}_{m,i}(x,y) \right], \tag{8}$$

where $I_{d,c}$ is the set of stimulus pairs in task d and condition c, $N_{d,c} = |I_{d,c}|$, and $s_{m,\text{task}}$ is the sample standard deviation of $u_{m,i}$ across all pairs in task d, as defined in Section 3.3 ($N = 280$ for Thatcher, $N = 288$ for Illumination, using $N - 1$ degrees of freedom). Here, $L^{\text{mod}}_{m,i}(x,y)$ and $L^{\text{orig}}_{m,i}(x,y)$ denote the spatial loss maps for the modified and original images of pair i, respectively (averaged across timesteps and noise realizations). For visualization in the figures, maps are cropped to remove video padding and resized to a common 96×96 square grid by bilinear interpolation, retaining each stimulus's presented orientation. All panels share a common symmetric color scale (the 99.5th percentile of absolute display-cell values across all 125 maps, ± 50.69), where warm colors denote loss increases, cool colors denote loss decreases, and white indicates zero change.



Figure S25: **Mean loss-change maps for image models (Part 1 of 2).**



Figure S26: **Mean loss-change maps for image models (Part 2 of 2).**

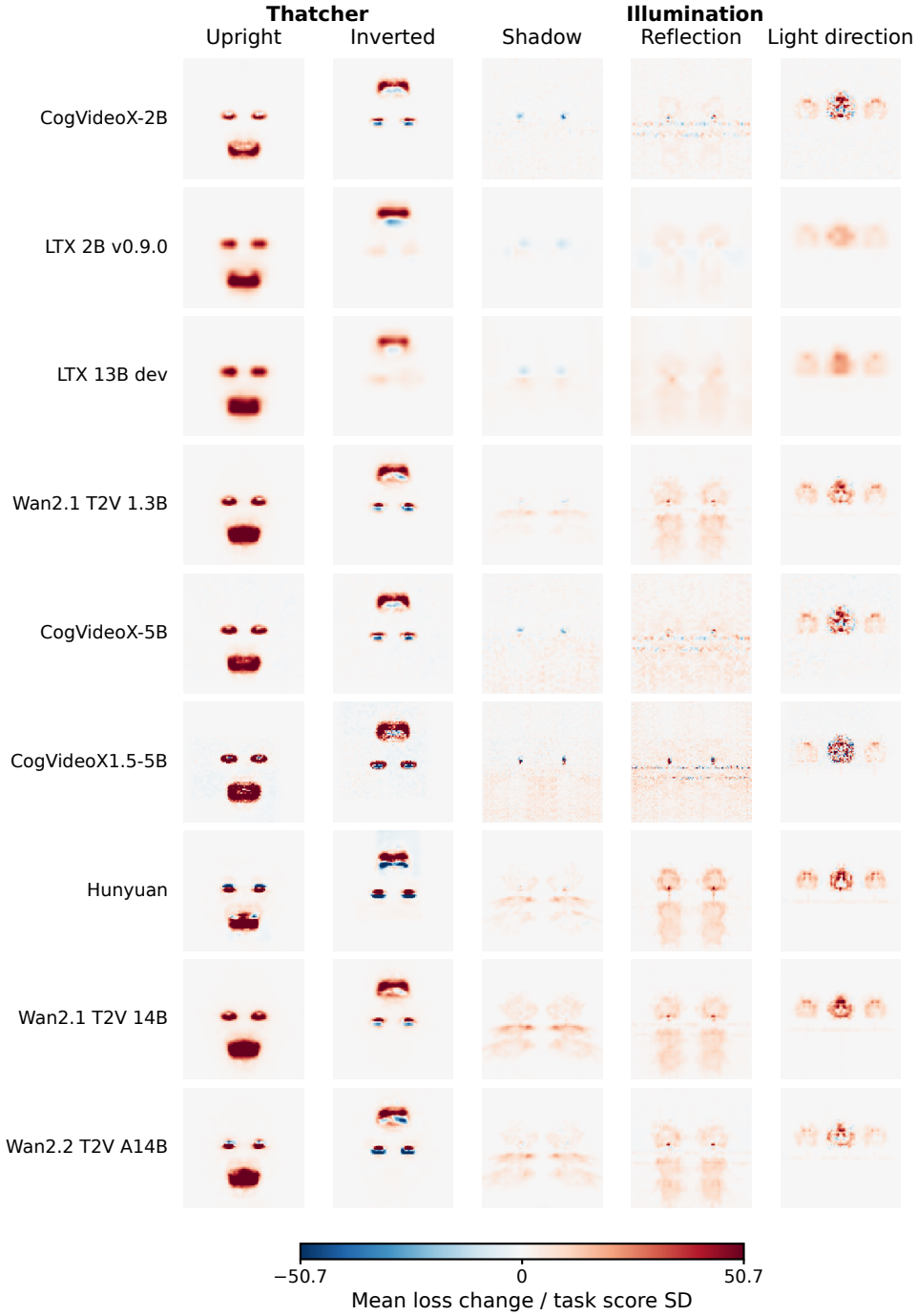

Figure S27: **Mean loss-change maps for video models.**

N TIMESTEP PROFILES AND PEAK LOCATIONS

In the main text (Section 4, Figure 4), we observed that sensitivity tended to peak later than human perceptual alignment along diffusion denoising schedules. This section provides the underlying evaluation details and complete model-wise data: (1) formal definitions of the profile metrics and schedule coordinates, (2) condition-specific peak comparisons across all 25 models (Figure S28), and (3) complete profiles for all evaluated models across conditions (Figures S29 and S30).

Metric definitions and schedule coordinates. To compare trajectories across heterogeneous model architectures and noise schedulers, we define:

- **Normalized schedule coordinate:** For a model with $|T|$ stored timesteps, discrete index $j \in \{0, \ldots, |T|-1\}$ is mapped to $t_{\text{norm}} = j/(|T|-1) \in [0, 1]$, where 0 denotes the earliest (maximum noise) step and 1 the final (minimum noise) step.
- **Standardized sensitivity:** For condition c at timestep t, sensitivity is the mean loss difference ($L_{\text{modified}} - L_{\text{original}}$) divided by the sample SD ($n-1$ denominator) pooled over all items in the task at that timestep. Sharing this task-wide denominator preserves relative response magnitudes across conditions.
- **Human alignment:** Item-level Pearson correlation r between model loss differences and human naturalness ratings, computed either pooled across the task or within each condition.
- **Peak locations:** Peaks are defined as signed maxima on each model's native discrete grid without smoothing or interpolation (exact ties break by the first stored index).

Condition-specific peak separation. Figure S28 presents within-model peak positions evaluated separately for each experimental condition. The extent of peak separation varied across interventions:

- **Thatcher task:** Sensitivity peaked later than alignment in 20 of 25 models for upright faces (median shift $\Delta = +0.184$) and 18 of 25 for inverted faces (median $\Delta = +0.192$).
- **Illumination task:** Sensitivity peaked later in 21 of 25 models for light direction ($\Delta = +0.204$), 22 of 25 for reflection ($\Delta = +0.576$), and 15 of 25 for cast shadow ($\Delta = +0.051$).

These condition-level breakdowns show that sensitivity generally peaks later than alignment across visual manipulations. The magnitude of this shift varies across physical cues: it is largest for surface reflections and lighting direction, and smaller for cast shadows.

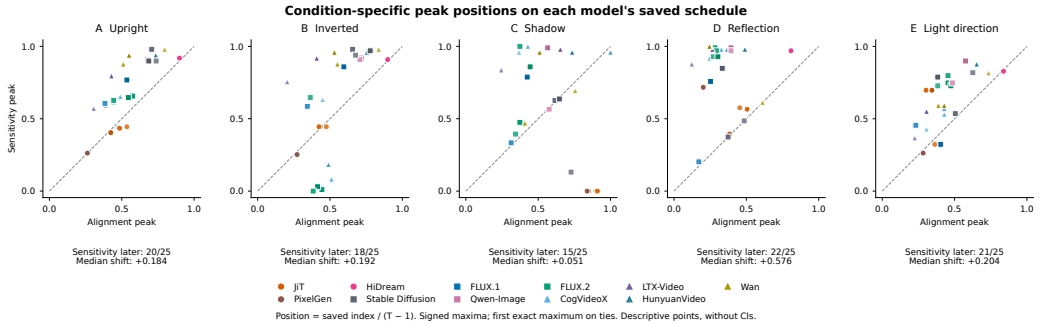

Figure S28: **Condition-specific alignment and sensitivity peak positions.** Each panel displays all 25 models (circles: pixel-native image models; squares: latent image models; triangles: video models; colored as in Figure 3). Horizontal and vertical axes show condition-specific alignment and sensitivity peak positions, respectively, on each model's normalized schedule ($[0, 1]$). Points above the diagonal indicate a later sensitivity peak (counts printed below each panel).

Individual model profiles. Figures S29 and S30 display the complete condition-specific alignment (solid curves, left axis) and sensitivity (dashed curves, right axis) trajectories for each of the 25 models. Raw axis limits are held constant across all panels within each task, with vertical origins aligned at zero without within-model rescaling. Filled and open circles mark the discrete alignment and sensitivity maxima, respectively, matching the coordinates plotted in Figure S28.

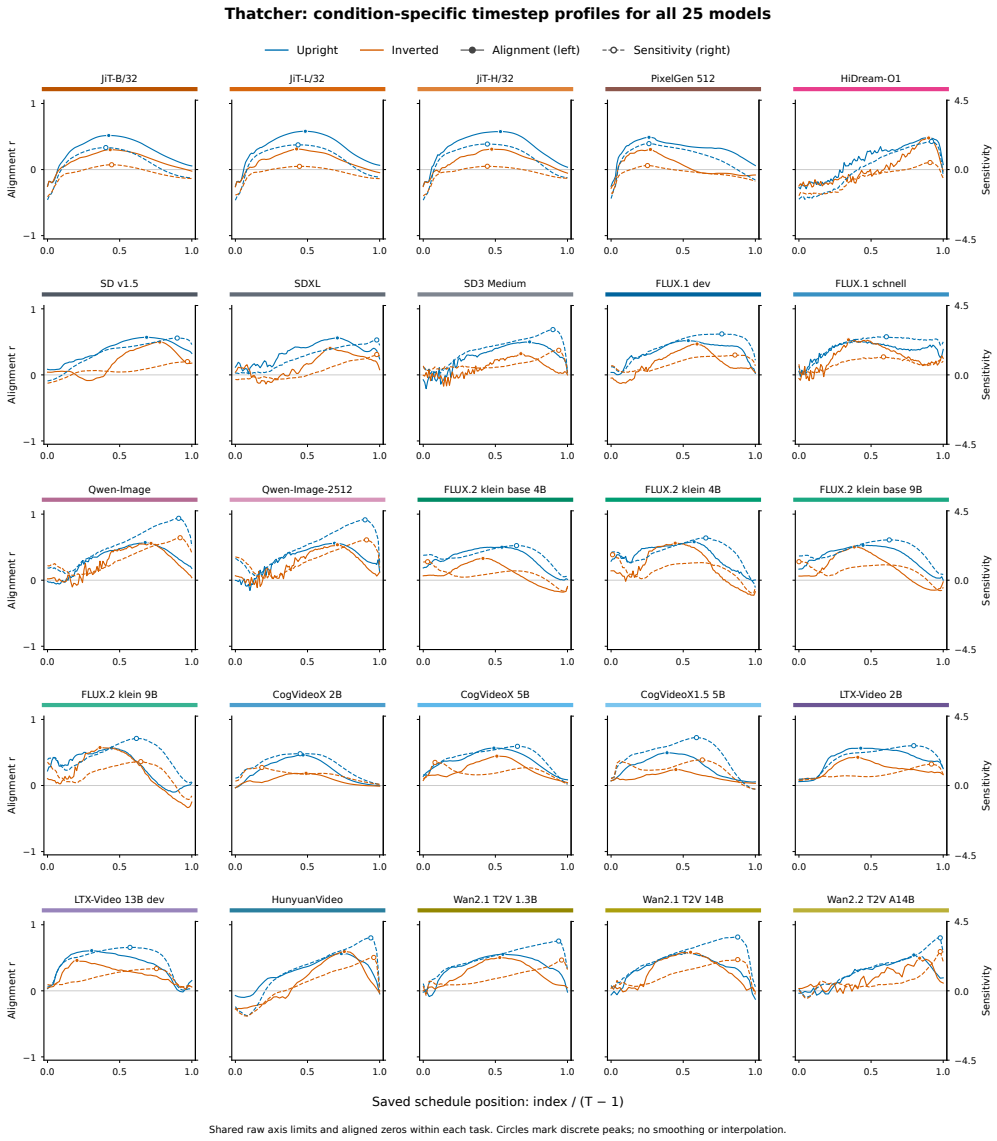

Figure S29: **Condition-specific Thatcher timestep profiles for all 25 models.** Panels show individual models (top bar colored as in Figure 3). Blue and orange curves denote upright and inverted faces; solid curves show item-level Pearson alignment (left axis) and dashed curves show task-scaled sensitivity (right axis). Filled and open circles mark alignment and sensitivity maxima, respectively. The horizontal axis indicates normalized schedule position ($[0, 1]$).

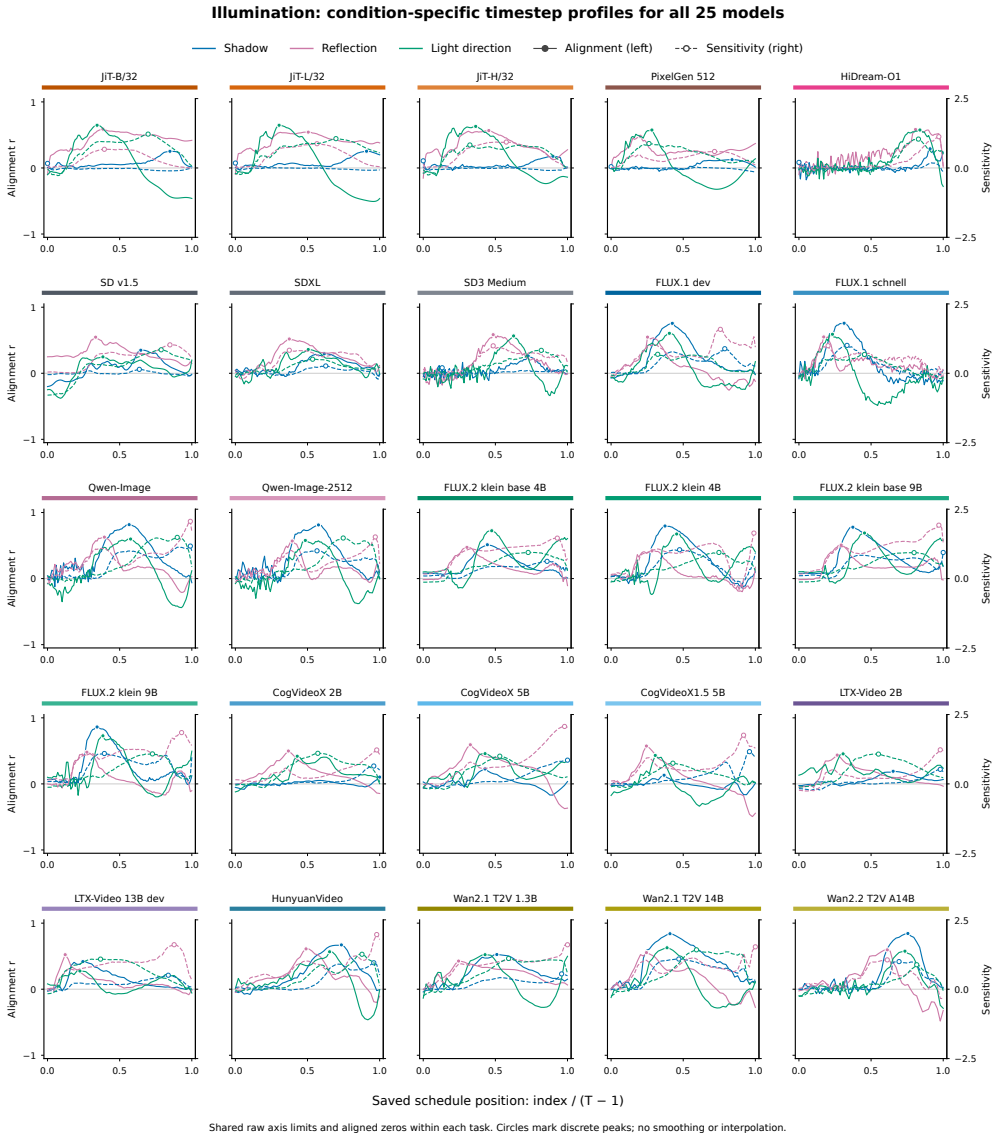

Figure S30: **Condition-specific Illumination timestep profiles for all 25 models.** Panels show individual models (top bar colored as in Figure 3). Blue, purple, and green curves denote shadow, reflection, and light-direction interventions; solid curves show item-level Pearson alignment (left axis) and dashed curves show task-scaled sensitivity (right axis). Filled and open circles mark alignment and sensitivity maxima, respectively. The horizontal axis indicates normalized schedule position ($[0, 1]$).

O CONTROL FOR TEXT CONDITIONING

To test whether our findings depend on empty text conditioning, we compared empty prompts with fixed descriptions of the stimulus content in a representative nine-model panel spanning diverse architectures (Stable Diffusion v1.5, SDXL, SD3, FLUX.1 dev, PixelGen, Qwen-Image, Wan2.1 T2V 1.3B, CogVideoX-2B, and LTX-Video 2B v0.9.0). Prompts described stimulus content (face orientation or a tabletop scene) without mentioning the visual manipulation or judging naturalness, and were identical for both the original and modified images within each pair (Table S7). Both prompt conditions were evaluated without classifier-free guidance, using the same timestep grids and noise draws (20 timesteps for image models, 10 for video models; five draws per timestep).

Table S7: **Fixed descriptive prompts for text conditioning.** Prompts used in the conditional evaluation. Each prompt was identical for the original and modified images within each pair and did not mention the visual manipulation or judge naturalness. The baseline condition used an empty string.

Condition	Image models	Video models
Thatcher (upright)	A close-up frontal photograph of a person's face, shown upright.	A static camera shot showing a close-up frontal view of a person's face, shown upright.
Thatcher (inverted)	An upside-down close-up frontal photograph of a person's face.	A static camera shot showing an upside-down close-up frontal view of a person's face.
Illumination (all conditions)	A computer-generated 3D scene of multiple geometric objects on a flat tabletop against a dark background.	A static camera shot of a computer-generated 3D scene with multiple geometric objects on a flat tabletop against a dark background.

Changes in human alignment (conditional minus empty Pearson r) ranged from $-.034$ to $.014$ on Thatcher and from $-.021$ to $.041$ on Illumination; Table S8 reports the paired estimates and 95% bootstrap confidence intervals for each model. Because the Thatcher prompts specified face orientation, we also assessed alignment within conditions: after removing orientation-specific means, correlations remained positive for all nine models under both prompt conditions. Across these nine models, human-alignment rankings had Spearman correlations of .967 on Thatcher and .983 on Illumination between prompt conditions. The sensitivity–alignment correlation changed from .659 to .594 on Thatcher and from .947 to .954 on Illumination. These results suggest that the observed differences in human alignment across models are largely preserved when simple descriptions of the stimulus content replace empty prompts.

Table S8: **Paired text-conditioning control.** Human alignment (Pearson r) under empty and descriptive (Desc.) prompts, with $\Delta r = r_{\text{Desc.}} - r_{\text{Empty}}$. Brackets give paired percentile 95% bootstrap confidence intervals for Δr from 10,000 participant-and-scene draws (140 face identities in Thatcher or the 48 base configurations within each operation in Illumination), shared between empty and descriptive prompts.

	Thatcher			Illumination		
Model	Empty	Desc.	Δr [95% CI]	Empty	Desc.	Δr [95% CI]
SD v1.5	.825	.822	$-.003$ $[-.006, .000]$	.234	.268	$+.033$ $[.017, .050]$
SDXL	.660	.662	$+.002$ $[-.028, .033]$	.226	.233	$+.007$ $[-.011, .025]$
SD3	.730	.744	$+.014$ $[-.002, .031]$	.263	.267	$+.004$ $[-.034, .044]$
FLUX.1 dev	.751	.752	$+.001$ $[-.007, .009]$	.353	.364	$+.011$ $[-.012, .031]$
PixelGen	.485	.493	$+.008$ $[-.044, .060]$	.101	.117	$+.016$ $[-.067, .091]$
Qwen-Image	.743	.709	$-.034$ $[-.049, -.018]$	.576	.555	$-.021$ $[-.060, .015]$
Wan2.1 T2V 1.3B	.784	.779	$-.004$ $[-.017, .008]$	.365	.406	$+.041$ $[.004, .083]$
CogVideoX-2B	.387	.379	$-.007$ $[-.014, -.001]$	.221	.215	$-.006$ $[-.044, .030]$
LTX-Video 2B	.753	.747	$-.006$ $[-.015, .003]$	.100	.109	$+.010$ $[-.031, .047]$

P CONTROL FOR SHARED SAMPLING NOISE

Directional sensitivity and human alignment are both functions of the same estimated paired scores. Computing both metrics from the same finite noise draws can therefore introduce correlated sampling errors, potentially inflating their cross-model association. Testing this control requires individual per-draw losses. While all 16 image models preserved per-draw losses across 100 timesteps in the primary evaluation, the primary 50-timestep video evaluation did not retain per-draw outputs. We therefore sampled three representative video models (Wan2.1 T2V 1.3B, CogVideoX-2B, and LTX-Video 2B v0.9.0) from a separate 10-timestep evaluation where per-draw losses were recorded, forming a combined panel of 19 models. Both subsets retained 20 draws and all stimulus pairs (280 Thatcher and 288 Illumination).

Within each model, we divided draws into fixed even/odd halves A and B of 10, using the same split across all items and timesteps, and recomputed sensitivity S and human alignment r in each half. The same-half estimate averages $\mathrm{corr}_m(S_A, r_A)$ and $\mathrm{corr}_m(S_B, r_B)$; the cross-half estimate averages $\mathrm{corr}_m(S_A, r_B)$ and $\mathrm{corr}_m(S_B, r_A)$. As a result, the same-half and cross-half associations differed by less than .001 in both tasks (Table S9), indicating that the relationship between directional sensitivity and human alignment is not driven by shared sampling noise.

Table S9: **Shared sampling noise control across 19 models.** Cross-model correlations between directional sensitivity and human alignment under same-half and cross-half draw splits, with $\Delta r = r_{\mathrm{Cross}} - r_{\mathrm{Same}}$.

Task	Same-half	Cross-half	Δr
Thatcher	.6962	.6971	$+.0009$
Illumination	.9242	.9238	$-.0005$