Revealing the Gap in Human and VLM Scene Description through Counterfactual Semantic Saliency

Ziqi Wen[1] **Parsa Madinei**[1,2] **Miguel P. Eckstein**[1,2]
[1]Department of Computer Science, University of California, Santa Barbara
[2]Department of Psychological and Brain Sciences, University of California, Santa Barbara
{ziqiwen, madinei, migueleckstein}@ucsb.edu

Abstract

Evaluating whether large vision-language models (VLMs) align with human perception for high-level semantic scene comprehension remains a challenge. Traditional white-box interpretability methods are inapplicable to closed-source architectures and passive metrics fail to isolate causal features. We introduce Counterfactual Semantic Saliency (CSS). This black-box, model-agnostic framework quantifies the importance of objects by measuring the semantic shift induced by their causal ablation from a scene. To evaluate AI-human semantic alignment, we tested prominent VLMs against a human psychophysics baseline comprising 16,603 valid responses across 307 complex natural scenes and 1,306 high-fidelity counterfactual variants. Our analysis reveals a pervasive scene comprehension gap: models exhibit an over-reliance (relative to humans) on object size, object centrality, and low-level saliency of objects. In contrast, models are less sensitive to the presence of people in the scenes than our human participants when describing the images. The over-reliance on object size is a primary driver explaining variations in model-human semantic divergence. Code and data are available at https://github.com/starsky77/Counterfactual-Semantic-Saliency-Data.

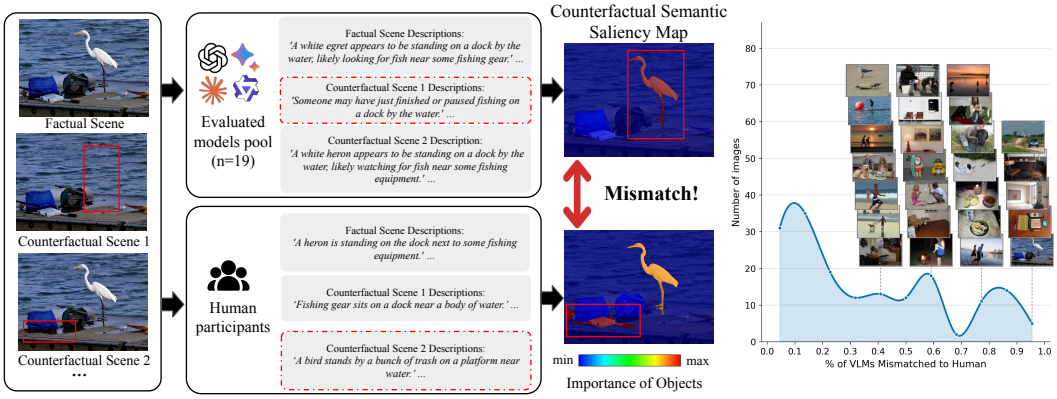


Figure 1: **Counterfactual Semantic Saliency (CSS) reveals a pervasive scene comprehension gap between humans and vision-language models (VLMs).** CSS quantifies the importance of an object by ablating it from a factual scene and measuring the semantic shift in the resulting descriptions. The object causing the maximal semantic shift is considered the most critical to scene perception. The example illustrates a mismatch between ALL 19 models and humans. The rightmost plot illustrates the percentage of the evaluated VLMs that failed to identify the human-consensus critical object for a given scene.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

As deep neural networks (DNNs) achieve superior performance, verifying whether these models exhibit perceptual strategies aligned with humans has become a critical challenge. Studies in AI-human alignment have revealed a persistent divergence in these strategies: even when achieving comparable accuracy on benchmark tasks, DNNs and human observers often rely on different features. Compared with humans, models tend to show over-reliance on local textures [21, 19], subtle pixel patterns [27], other localized cues that are highly inconsistent with holistic human perception [16, 62], or even ignoring the spatial relationship of visual elements[65]. While scaling parameters and training data have been shown to induce emergent alignment [20, 22, 36], these evaluations have primarily focused on isolated, simplified tasks such as basic object classification. Consequently, understanding the gap between modern Vision-Language Models (VLMs) and human observers in the context of scene comprehension, which requires integrating multiple objects, spatial contexts, and complex semantic relationships, remains largely underexplored.

To investigate this alignment gap, the driving factors behind model decisions need to be understood. However, the scale, complexity, and frequently closed-source nature of modern VLMs have made many traditional "white-box" interpretability techniques impractical. Evaluating the mechanisms of these architectures requires a shift toward black-box, model-agnostic strategies. Causal intervention frameworks [59, 24, 53, 14] have emerged as a powerful solution; by systematically manipulating input variables, such as ablating specific visual elements, we can directly measure the downstream effects and isolate the causal weight of individual scene components.

Building upon this paradigm, we propose Counterfactual Semantic Saliency (CSS), a novel, black-box, and model-agnostic methodology designed to evaluate object importance in scene perception. CSS enables us to systematically investigate *whether modern VLMs perceive and prioritize elements in complex scenes in alignment with human observers.* To compare model and human semantic alignment, we constructed a human baseline using complex, multi-object images from a widely adopted human saliency dataset [64]. Using state-of-the-art generative inpainting with Nano Banana 2 [57], we created 307 factual scenes and 1,306 high-fidelity counterfactual variants by systematically ablating 3–6 visual elements per scene. We then conducted a large-scale human psychophysics study, collecting 16,603 valid textual descriptions across all 1,613 images from 227 independent participants.

Pervasive gaps between models and human scene comprehension were found through statistical analysis. Notably, every evaluated VLM failed to reach human-human consistency levels when predicting the most critical object for scene comprehension or the object importance ranking. Furthermore, we identified divergences in the underlying attribute sensitivity during scene viewing: whereas humans focus more on the presence of people, models heavily rely on bottom-up attributes, including object size, object centrality, and low-level saliency. Among these divergent factors, we demonstrate that exaggerated VLM sensitivity to object size is the dominant factor contributing to the human-AI semantic alignment gap.

The primary contributions of this work are as follows:

- We introduce Counterfactual Semantic Saliency (CSS), a black-box and model-agnostic methodology to systematically quantify the importance of objects in scene description, overcoming the limitations of traditional white-box interpretability methods.
- We establish a human baseline by conducting a large-scale human psychophysics study, yielding a comprehensive benchmark dataset comprising 16,603 valid responses across 307 factual scenes and 1,306 counterfactual variants from 227 independent participants.
- We provide the first systematic quantification of the semantic alignment gap between modern VLMs and human observers in scene description, revealing a pervasive divergence in the sensitivity to 4 object attributes (size, centrality, low-level saliency, person-non-person) between VLMs and humans.
- Through robust statistical analysis, we discover the major driving factor of this scene comprehension gap, demonstrating that an excessive reliance on size in modern architectures contributes the most to their divergence from humans.

2 Related Works

Historically, understanding the decision-making processes of visual models like CNNs relied heavily on white-box methods, ranging from gradient-based [51] and activation-based [4] internal analyses to early black-box probing [47, 41]. As Vision Transformers become the dominant vision encoder architecture, visualizing attention maps provides cues about the features that contribute to a model's output [9, 35, 30]. However, the immense complexity of modern VLMs makes these straightforward techniques no longer sufficient to fully capture the true causal features driving a model's generation [29, 10]. While black-box analytical methods have long existed, previous approaches to complex visual tasks have faced significant obstacles. When altering image inputs for causal intervention, researchers either use an oversimplified method that blocks different scene regions [43], or, when altering the scene, need to ensure that the modified image quality does not introduce confounding artifacts [53], or they have to collect naturally occurring image pairs that were visually similar but semantically distinct [58]. Although traditional inpainting models could rapidly remove objects from scenes [37, 49, 54, 69], the visible traces left behind by these modifications were non-negligible.

However, recent advancements in generative AI have made high-quality, flawless, and fine-grained image editing viable [45, 57, 60, 63]. SOTA closed-source models are capable of generating images that are nearly indistinguishable to human observers, and the maturation of image-to-image generation has enabled highly controllable editing [18, 50, 31]. This technological leap serves as the foundational premise for our work. Counterfactual Semantic Saliency (CSS) leverages the most advanced generative models available (Nano Banana 2) to produce high-fidelity counterfactual images, allowing us to systematically verify whether a model's semantic shifts align with human perception. Crucially, the exceptional visual quality ensures that neither the evaluated models nor the human participants detected generative artifacts, providing the necessary foundation for an unbiased comparison.

3 Method

3.1 Counterfactual Semantic Saliency (CSS)

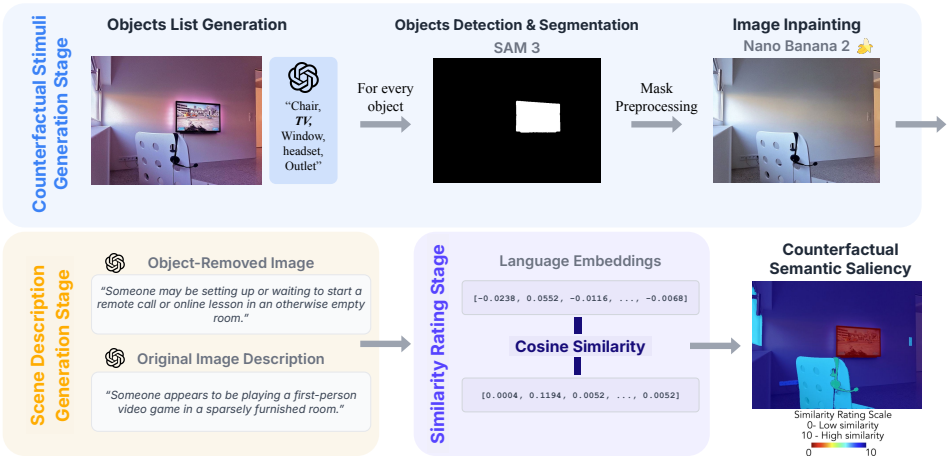


Figure 2: **Workflow of Counterfactual Semantic Saliency.** (1) Counterfactual Stimuli Generation Stage: A large vision-language model generates a list of object names in the scene. For each identified object, SAM3 generates a corresponding segmentation. Following mask preprocessing, an image inpainting model (Nano Banana 2) removes the target object. In the illustration, the TV, along with the pink LED light, is removed from the scene. (2) Scene Description Stage: Descriptions are sampled from the target evaluation model for both the original image and all counterfactual variants. (3) Similarity Rating Stage: Counterfactual Semantic Saliency is quantified as one minus the mean pairwise cosine similarity between the sentence embeddings of the factual scene and counterfactual scene descriptions.

3.1.1 Counterfactual Stimuli Generation

Counterfactual Semantic Saliency (CSS) defines object importance by measuring the semantic shift induced by its removal as a counterfactual intervention. To generate these counterfactual stimuli, the initial stage consists of the following three steps: object list generation, object segmentation, and image inpainting, as illustrated in Fig.2.

Object list generation: We applied GPT4o[1] to generate a list of object names given the scene. Objects that could be considered part of the background (e.g., sky, sea, wall, ground) are explicitly excluded; the removal of those objects would induce disproportionate semantic shifts or obvious hallucination due to their spatial dominance. We generated N distinct responses via multinomial sampling and merged them using GPT-4o to eliminate duplicated labels. Detailed prompts are provided in the Appendix.

Object Segmentation & Mask Preprocessing: Segmentation masks were generated using the text-prompted segmentation model, SAM3 [6], with a confidence threshold of 0.4. To isolate meaningful visual elements before object removal, we applied a four-step preprocessing pipeline. First, duplicated masks are removed by checking all possible pairs of masks whose Intersection over Union (IoU) is greater than 95%. Second, guided by principles of perceptual grouping [44], we merged spatially proximate masks (distance < 30 pixels) that have identical labels into a singular mask. This ensures that higher-order structural elements (e.g., clustered candles forming a heart shape) are treated as unified conceptual entities rather than disjointed features, and individual masks were discarded after merging. Third, we discarded masks occupying >30% of the total image area to prevent severe hallucination and context collapse associated with removing large structural components. Finally, a dilation was applied to ensure complete target coverage before the inpainting process.

Image Inpainting: While conventional inpainting models such as MAT [37] can remove the object effectively and efficiently, they often introduce artifacts when handling large masks or complex occlusions. Therefore, we utilized the SOTA generative model Nano Banana 2 [57] to generate high-fidelity, artifact-free counterfactual images. To control for model hallucinations or any undefined behavior introduced by the model, all generated stimuli underwent rigorous human validation. Fourteen independent annotators (recruited via Prolific) verified the successful ablation of the target objects. Images flagged for incomplete removal were manually reviewed and excluded from the final experimental dataset. Detailed validation protocols are provided in the Appendix.

3.1.2 Measuring Semantic Shift with CSS scores

We formalize Counterfactual Semantic Saliency (CSS) through a causal intervention framework. Let $\mathcal{I}$ denote the original scene, o the target objects, and $\mathcal{M}$ the target vision-language model evaluated for scene description. We represent the targeted ablation of o using *do*-operator as the intervention $do(o = \emptyset)$, which yields the counterfactual stimulus $\mathcal{I}_{do(o=\emptyset)}$. To account for the variability in the model's generated descriptions, we sample a set of N descriptions (denoted by d) for both the factual condition, $D = \{d_1, \ldots, d_N\} \sim \mathcal{M}(\mathcal{I})$, and the counterfactual condition, $D' = \{d'_1, \ldots, d'_N\} \sim \mathcal{M}(\mathcal{I}_{do(o=\emptyset)})$. Let $\mathcal{E}(\cdot)$ denote the sentence embedding function, which we implement using the open-source sentence embedding model Jasper [66]. The CSS score of object o is defined as one minus the expected cosine similarity between the semantic representations of the factual and counterfactual descriptions. We approximate this semantic shift as one minus the mean pairwise cosine similarity between the two sampled sets:

$$\text{CSS}(o) = 1 - \frac{1}{N^2} \sum_{j=1}^{N} \sum_{k=1}^{N} \frac{\mathcal{E}(d_j) \cdot \mathcal{E}(d'_k)}{\|\mathcal{E}(d_j)\| \|\mathcal{E}(d'_k)\|}$$

where $N = 5$ for all the models. Under this formulation, CSS score acts as a metric of perceptual importance. For each scene, the semantically critical object is defined as the object that, when removed, impacts the scene description the most (highest CSS score). Finally, these object-level CSS scores are mapped back onto their corresponding segmentation masks to construct a top-down counterfactual semantic saliency map for the entire visual scene.

3.2 Dataset

We evaluated our framework using the Object and Semantic Images and Eye-tracking (OSIE) dataset [64], a standard benchmark in visual saliency research, consisting of 700 complex natural scenes with

paired free-viewing eye-tracking data [12]. The rich, multi-object composition of these foregrounds provides an ideal testbed for targeted counterfactual analysis, enabling the systematic isolation of semantic contributions. To maintain a controlled density of semantic elements and prevent overly cluttered or sparse interventions, we filtered the initial pool of images to include only scenes containing between 3 and 6 counterfactual variants (98 three-variant scenes, 87 four-variant scenes, 68 five-variant scenes, and 54 six-variant scenes). Following the generation of counterfactual stimuli and human validation of the inpainting quality, our final dataset contains 307 unique factual scenes and 1,306 counterfactual variants.

3.3 Human Psychophysics

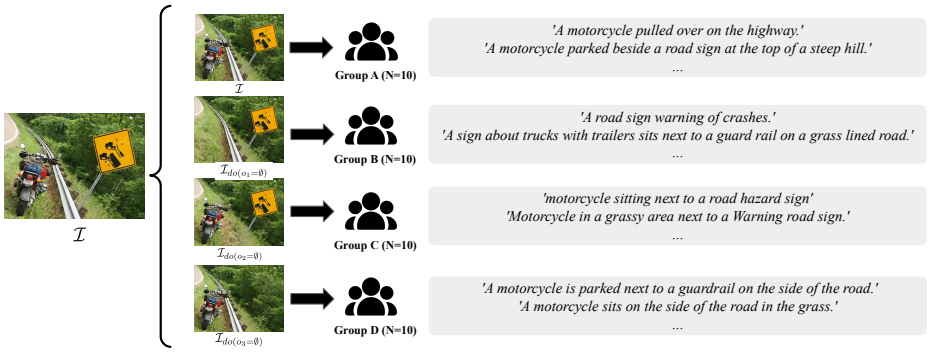


Figure 3: **Between-subjects experimental design for human psychophysics.** A scene containing N target objects (1 factual image $\mathcal{I}$ and N counterfactual variants $\mathcal{I}_{do(o_i=\emptyset)}$) was distributed across $N+1$ mutually exclusive stimulus sets. Each set was evaluated by an independent group of 10 participants, ensuring no individual was exposed to multiple variants of the same scene. We recruited 227 Subjects to complete a total of 16,665 trials.

To evaluate the alignment between vision-language models and human observers in scene comprehension, we established a human baseline for Counterfactual Semantic Saliency, with 227 subjects and 16,665 trials. Mirroring the pipeline used to evaluate the models, we collected human scene descriptions for all factual scenes and their corresponding counterfactual variants via the Prolific platform.

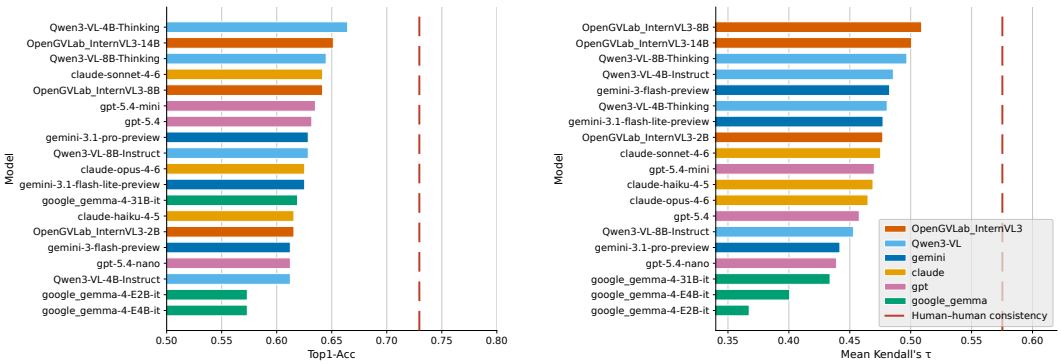
To prevent memory and carryover effects, it was strictly required that no single participant view multiple variants of the same scene. We therefore implemented a between-subjects experimental design, as illustrated in Fig.3. Factual scenes were first divided into four groups based on their target object count (ranging from 3 to 6 objects). For a scene containing *N* target objects, the factual image and its *N* counterfactual variants were distributed across *N* + 1 mutually exclusive stimulus sets. Each stimulus set was then assigned to an independent group of 10 participants.[1] This routing protocol ensured that we collected 10 independent textual descriptions for each unique stimulus, while strictly guaranteeing that no participant was exposed to more than one condition for any given scene. To filter low-quality responses that are oversimplified or irrelevant to the scene, we discard responses whose mean cosine similarity of sentence embeddings [66] with other human responses is below 0.4. Of the 16,665 trials, 62 (~0.3%) are discarded in this process.

4 Results

4.1 Gap between Human and Model Scene Comprehension

We quantified the alignment between human observers and all the commonly used open-source [68, 33, 3] and closed-source [52, 2, 11] vision-language models (n=19) using two primary metrics: Top-1 Accuracy and Kendall's rank correlation coefficient (τ). Top-1 Accuracy measures the strict match rate, defined as the proportion of scenes where the model correctly identified the single most

[1] Incidental oversampling occurred for certain stimuli during crowdsourcing; all collected descriptions were kept in the final dataset.



(a) Top-1 Accuracy at predicting the critical object.

(b) Mean Kendall's τ between models and humans.

Figure 4: **Model-human alignment on scene description. (a)** Top-1 Accuracy representing the proportion of scenes where the evaluated model successfully identified the most semantically critical object, as determined by humans. **(b)** Mean Kendall's rank correlation coefficient (τ) across all scenes, capturing the overall ordinal alignment of object importance hierarchies between models and humans. Full distributions of Kendall's τ for each model are provided in the Appendix. All the gaps here are significant with a one-tail bootstrap test ($p < .01$).

critical object (i.e., the target whose ablation induced the highest CSS score) as determined by the human baseline. Furthermore, we employed Kendall's τ to evaluate the structural alignment, measuring the ordinal correlation between the full rank-ordered hierarchies of object importance generated by the models versus humans.

To estimate human-human consistency, we partitioned the human data from 10 participants per counterfactual scene into two groups of five participant descriptions: one reference group and one comparison group. For the reference group, descriptions for counterfactual scenes from 5 participants were compared to 10 descriptions from factual scenes using cosine similarity. To mitigate individual subjective variance, we computed a human aggregate metric of CSS score, averaged across 50 description pairs' cosine similarity (5 counterfactual vs. 10 factual). All VLMs were subsequently evaluated against the reference group to ensure a controlled comparison. For human–human consistency baseline, we computed CSS maps from the comparison group and compared them with the CSS maps from the reference group using Top-1 accuracy and Kendall's τ [2]. As illustrated in Fig. 4, a persistent and systematic gap exists between human and model performance in counterfactual scene comprehension. While the human-human consistency for Top-1 Accuracy reached 73%, all evaluated models fell significantly short, ranging from 57% to 65%. This divergence is similarly reflected in the structural alignment metric: the human-human consistency achieved a mean Kendall's τ of 0.58, whereas model scores ranged from 0.37 to 0.51. Importantly, this divergence cannot be attributed to the choice of sentence embedding model, artifacts introduced by the inpainting model, or the variability of human agreement. We present further ablations in the Appendix confirming the robustness of this scene comprehension gap.

4.2 Attribute Sensitivity in Models and Humans

4.2.1 Possible Hypotheses

To identify the underlying factors driving the semantic shift resulting from object removal, we draw on established literature on human visual attention and DNN model bias. We investigated four primary attributes that are believed to influence CSS: size, centrality, low-level saliency, and whether the object is a person. We measured how these attributes influence object importance by correlating the CSS of removed objects with the objects' attributes across all 1,306 counterfactual scenes, referred to as "sensitivity to X" in the following. Spearman correlation is applied for numerical attributes (e.g., object size), and point-biserial correlation is applied for binary attributes (e.g., Person-Non-person).

[2]Due to the incidental oversampling and filtering described previously, this comparison group occasionally contained between four and seven responses rather than five.

For the following analysis, all of the collected subjects' responses (10 subjects for each counterfactual scene) are used to calculate human CSS maps, referred to as the human consensus.

Sensitivity to Size: Physical stimuli size acts as a primary bottom-up cue for visual attention in humans in early psychology research[17, 7]. However, deep neural networks exhibit an exaggerated, unnatural reliance on spatial dominance during feature extraction that is inconsistent with human perception, especially for object detection [15, 26, 61]. We hypothesize that this disproportionate reliance on size persists even within state-of-the-art VLMs, likely inherited from their foundational CNN or Vision Transformer architectures. We quantified the size of the object using the area of the masks.

Sensitivity to Center: Human free-viewing tends to fixate on the center of a scene [56]. Meanwhile, because large-scale training datasets (e.g., ImageNet [13], MS-COCO [39]) are heavily saturated with centrally framed objects, deep learning models could inherit and amplify this spatial prior, biasing their semantic interpretation to the center of the image regardless of contextual relevance. We used the negative of the distance between the center of the masks and the center of the image to measure the centeredness for each object.

Sensitivity to Low-level Saliency: Early theories of visual attention have shown that bottom-up, low-level image statistics, such as local contrasts in luminance, color, and orientation, automatically capture human gaze [28]. To measure this low-level signal without introducing the confounding semantic variables present in modern neural networks, we quantified the low-level saliency of each ablated object using the Graph-Based Visual Saliency (GBVS) algorithm [25]. Specifically, the low-level visual saliency of each ablated object was quantified as the maximum GBVS activation value located within its corresponding segmentation mask.

Sensitivity to Person/Non-person: Research in human eye movements demonstrates that observers consistently prioritize faces and human figures during free-viewing, independent of explicit task demands [8, 32]. However, whether vision-language models intrinsically replicate this sensitivity during scene comprehension remains systematically underexplored. We utilized GPT-5.4 to classify each ablated object as either "person" or "non-person" based on its label during counterfactual stimuli generation.

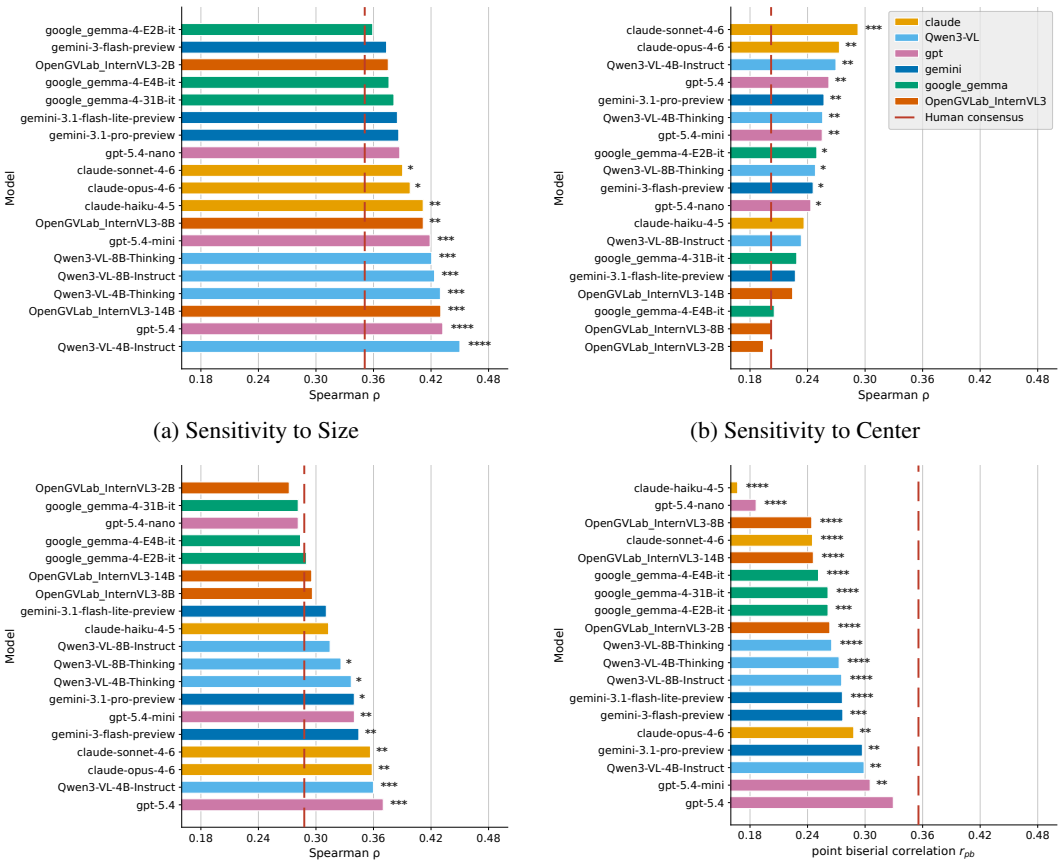
4.2.2 Divergence in Attribute Sensitivity

To evaluate the four hypothesized drivers of semantic shift, we quantified the specific attribute sensitivity of all evaluated models and compared them directly against the human consensus baseline. As illustrated in Fig.5, we measured the correlation between the target object attributes and the Counterfactual Semantic Saliency. To formally assess whether the models' reliance on these visual attributes significantly deviated from human perception, we conducted one-tailed bootstrap significance tests (n = 10,000) across the entire counterfactual image set.

Furthermore, to ensure the attribute sensitivity we observed from human consensus (dashed line in Fig.5) was not an artifact of random statistical noise, we used permutation tests (n = 10,000) across all object attribute labels (Size, Center, Low-level Saliency, and Person-Non-person). Correlating the permuted data with the human consensus under the null hypothesis yielded a mean approximating zero with minimal variance ($std \approx 0.027$) for all 4 factors, confirming that the attribute sensitivity of humans identified here is statistically significant ($p < .0001$). The comparative analysis reveals a divergence in how VLM and humans prioritize information:

- **Over-reliance on low-level attributes:** Deep learning models rely heavily on basic spatial and pixel-level features. Compared to human observers, models are significantly more sensitive to object size (Fig. 5a), object centrality (Fig. 5b), and low-level saliency (Fig. 5c). They are more frequently influenced by objects that are large, central, or highly contrasting, regardless of their actual semantic value.

- **Insensitivity to the presence of people:** Conversely, models lack the specialized prioritization that humans apply to the scene. While the description from human viewers is heavily influenced by the presence of people in the scene, the evaluated models are less sensitive to whether people are present or absent in the scene. (Fig. 5d).



(a) Sensitivity to Size

(b) Sensitivity to Center

(c) Sensitivity to Low-level Saliency

(d) Sensitivity to Person-Non-person

Figure 5: **Divergence of Attribute Sensitivity:** We computed the correlation between specific object attributes and the resulting CSS score upon ablation. Significance tests indicate the statistical deviation of each model's correlation coefficient from the human baseline ($* = p < .05$; $** = p < .01$; $*** = p < .001$; $**** = p < .0001$; derived from a one-tailed bootstrap test).

4.3 Driving Factors for Semantic Shift Gap

Having quantified the divergence of attribute sensitivity between VLMs and humans (Fig. 5), we need to determine which specific factors drive the observed disparity in critical object identification (Fig. 4a). To achieve this, we evaluated whether the magnitude of the sensitivity divergence (Δr) between models and humans for each specific attribute correlates with the overall performance gap in Top-1 Accuracy (ΔAcc) between evaluated models and humans.

To isolate statistically significant driving factors from baseline random effects, we used permutation tests. We performed 10,000 permutations across the target attribute labels, recalculated the corresponding Δr for the null distribution, and evaluated its correlation with ΔAcc. As shown in Table 1, while the null distributions are centered near zero, the variance indicates a cautious interpretation of the correlations. Among the four hypotheses, Sensitivity to Size emerged as the dominant driving factor of the gap between humans and VLMs ($r = -0.6130$). The excessive reliance on object size explains the systematic model failure illustrated in Fig. 1; because the white bird occupied a significantly larger pixel area than the fishing rod, the semantic shifts obscured the importance of the fishing rod, leading to misalignment with the human consensus.

4.4 The Mismatch Between White-Box Attention and Black-Box Causality

Attention and activation maps have historically served as diagnostic tools for visualizing the features that contribute to a neural network's response. However, it remains an open question whether these

Table 1: **Driving factors of the VLM-Human scene comprehension gap.** The correlation between the divergence in specific attribute sensitivity (Δr) and the overall gap in Top-1 Accuracy (ΔAcc) across evaluated models. The null distribution metrics were obtained via permutation tests to isolate random baseline effects. The p-values here were obtained by one-tailed permutation tests

Δr Type	**Corr. to ΔAcc**	**Mean of Null Dist.**	**Std of Null Dist.**	**P-value**
Size	-0.6130	0.0023	0.3193	0.0211
Center	-0.2296	-0.0030	0.3189	0.2635
Low-level Saliency	-0.3746	-0.0012	0.3155	0.1273
Person-Non-person	-0.1533	-0.0014	0.3247	0.3398



Figure 6: Comparison between CSS Maps and attention-based white-box maps

white-box observational methods can yield interpretable, object-level importance in large VLMs comparable to our causal black-box framework. To compare CSS against traditional interpretability techniques, we applied three widely adopted methods, raw Attention maps, Grad-CAM[51], and Attention-Guided CAM (AGCAM)[35], to an open-source VLM (Qwen3-VL-4B-Instruct).

Because VLM text generation is autoregressive, these white-box methods yield distinct spatial saliency maps (not to be confused with low-level saliency maps such as GBVS) for every generated token. To capture the global influence of the input pixels on the entire generated description, we aggregate all the saliency maps generated for each token through a max function. Finally, to enable a direct comparison with our object-level CSS metric, the white-box saliency of each visual element is formally quantified as the maximum activation value located within its corresponding segmentation mask.

Qualitative comparisons between these raw white-box saliency maps (before object masking) and CSS are illustrated in Fig. 6. Notably, the white-box maps derived from token-image activations severely lack human interpretability; they are frequently either incomprehensibly cluttered or overly sparse, and often focus on semantically irrelevant spatial regions. Consequently, the correlation between these traditional object-level saliency scores and the CSS metric is low and negative (mean Kendall's τ across all the images with the CSS: raw attention: -0.3352, GradCAM: -0.2998, AGCAM: -0.2420). This mismatch reveals the limitations of passive interpretability techniques when applied to the complex architectures of modern VLMs. More comparisons between our methods and white-box saliency maps are provided in the Appendix.

4.5 Vision-Only Models' Observational Saliency Fails to Capture CSS

Salient Object Detection (SOD) models [46, 40] are trained on large-scale datasets of human-annotated salient object masks, providing a vision-only estimate of which objects humans perceive as critical in a scene. However, similar to the white-box methods discussed in Sec. 4.4, SOD remains an observational approach: it infers object importance passively from image appearance without intervening on the scene. We therefore ask whether such a vision-only method can predict the object importance revealed by CSS.

Here, we evaluated two commonly used SOD models, U^2-Net [46] and Visual Saliency Transformer (VST) [40], on the same 307 original images used in our evaluation. Following a similar procedure to that used for the attention-based white-box maps in Sec.4.4, we used the maximum and mean values of the U^2-Net /VST saliency map within the corresponding segmentation mask in CSS to convert the pixel-level saliency maps into object-level saliency maps. Those SOD maps are treated the same way as the VLMs' CSS maps, and we repeat the same evaluation reported in Fig.4. The performance of

U2Net and VST using both maximum and mean values is shown in Table.2. The performance of Top1-Acc from both SOD models falls below all the VLMs we evaluated, indicating that vision-only approaches, even when trained on human-annotated data, struggle to capture the same object-level importance as CSS.

Table 2: Performance of Salient Object Detection on predicting the critical object and rank order of object importance. gemma-4-E4B represents the VLM with the worst Top-1 Acc performance.

Baseline	Mean $\tau \uparrow$	Top1-Acc $\uparrow$
VST (max) [40]	0.378	0.414
VST (mean) [40]	0.258	0.345
U^2-Net (max) [46]	0.353	0.476
U^2-Net(mean) [46]	0.229	0.384
gemma-4-E4B	0.401	0.573

5 Discussion

Data Distribution and Inductive Bias: The over-reliance on low-level attributes from the models we discovered is likely to originate from the "object-centered" nature of foundational vision datasets. Most vision encoders are pretrained with images in which the primary subject is centrally framed and spatially dominant [55], leading the encoder to mistake size for importance. Consequently, smaller or peripheral objects, which may be semantically indispensable to a human viewer, show only a minor influence on the model's internal representation. This dependence is further compounded during the vision-language alignment [19]. Most image-caption pairs sourced from the Internet typically focus on spatially dominant entities [23, 42]. As a result, even if the vision encoder captures a complete scene representation, the language output remains more sensitive to object size regardless of its actual semantic importance.

Superhuman Inference and Reasoning Gap: The "superhuman" capabilities of modern VLMs, specifically their advanced reasoning, can also exacerbate the human-AI mismatch. An example is illustrated in Fig.1: the ablation of the fishing rod fundamentally shifted the human interpretation of the scene (e.g., "fishing equipment" -> "a bunch of trash"). In contrast, models were often able to infer the ablated context from surrounding cues or, in some cases, exhibited hallucinatory behavior by describing the ablated object as if it were still present [48, 5, 38].

Limitation: While this work characterizes how the alignment gap between models and humans manifests in scene description, the underlying causes of this divergence remain unexplored. For example, the models' insensitivity to the presence or absence of a person in the scene could arise because they hallucinate people in scenes where none are present, or because they mention people less often than human observers do. Either case could produce the observed divergence. Identifying which factors are responsible will be crucial for developing vision-language models that are better aligned with human perception.

6 Conclusion

We introduced Counterfactual Semantic Saliency (CSS), a novel black-box framework designed to quantify the importance of objects in scene comprehension. By benchmarking modern VLMs against a large-scale human psychophysics baseline, we uncovered a pervasive alignment gap in scene comprehension. We demonstrate that while human observers naturally rely on the presence of people in the scene, modern VLMs default to an over-reliance on low-level visual features (object size, centrality, saliency).

7 Acknowledgment

Research was sponsored by the U.S. Army Research Office and accomplished under cooperative agreement W911NF-19-2-0026 for the Institute for Collaborative Biotechnologies.

References

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] A. Anthropic. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1(1):4, 2024.

[3] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[4] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549, 2017.

[5] A. F. Biten, L. Gómez, and D. Karatzas. Let there be a clock on the beach: Reducing object hallucination in image captioning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1381–1390, 2022.

[6] N. Carion, L. Gustafson, Y.-T. Hu, S. Debnath, R. Hu, D. Suris, C. Ryali, K. V. Alwala, H. Khedr, A. Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

[7] U. Castiello and C. Umiltà. Size of the attentional focus and efficiency of processing. *Acta psychologica*, 73(3):195–209, 1990.

[8] M. Cerf, E. P. Frady, and C. Koch. Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of vision*, 9(12):10–10, 2009.

[9] H. Chefer, S. Gur, and L. Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 397–406, 2021.

[10] H. Chefer, S. Gur, and L. Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 782–791, 2021.

[11] G. Comanici, E. Bieber, M. Schaekermann, I. Pasupat, N. Sachdeva, I. Dhillon, M. Blistein, O. Ram, D. Zhang, E. Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[12] B. De Haas, A. L. Iakovidis, D. S. Schwarzkopf, and K. R. Gegenfurtner. Individual differences in visual salience vary along semantic dimensions. *Proceedings of the National Academy of Sciences*, 116(24):11687–11692, 2019.

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[14] S. Ding, S. Vasa, and A. Ramadwar. Explanation-driven counterfactual testing for faithfulness in vision-language model explanations. *arXiv preprint arXiv:2510.00047*, 2025.

[15] M. P. Eckstein, K. Koehler, L. E. Welbourne, and E. Akbas. Humans, but not deep neural networks, often miss giant targets in scenes. *Current Biology*, 27(18):2827–2832, 2017.

[16] T. Fel, I. F. Rodriguez Rodriguez, D. Linsley, and T. Serre. Harmonizing the object recognition strategies of deep neural networks with humans. *Advances in neural information processing systems*, 35:9432–9446, 2022.

[17] P. M. Fitts. The information capacity of the human motor system in controlling the amplitude of movement. *Journal of experimental psychology*, 47(6):381, 1954.

[18] T.-J. Fu, W. Hu, X. Du, W. Y. Wang, Y. Yang, and Z. Gan. Guiding instruction-based image editing via multimodal large language models. *arXiv preprint arXiv:2309.17102*, 2023.

[19] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[20] R. Geirhos, K. Narayanappa, B. Mitzkus, T. Thieringer, M. Bethge, F. A. Wichmann, and W. Brendel. Partial success in closing the gap between human and machine vision. *Advances in Neural Information Processing Systems*, 34:23885–23899, 2021.

[21] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International conference on learning representations*, 2018.

[22] A. Gokce and M. Schrimpf. Scaling laws for task-optimized models of the primate visual ventral stream. *arXiv preprint arXiv:2411.05712*, 2024.

[23] J. Gordon and B. Van Durme. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 25–30, 2013.

[24] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, and S. Lee. Counterfactual visual explanations. In *International Conference on Machine Learning*, pages 2376–2384. PMLR, 2019.

[25] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. *Advances in neural information processing systems*, 19, 2006.

[26] T. R. Hayes and J. M. Henderson. Deep saliency models learn low-, mid-, and high-level features to predict scene attention. *Scientific reports*, 11(1):18434, 2021.

[27] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

[28] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259, 1998.

[29] S. Jain and B. C. Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, 2019.

[30] S. Jo, G. Jang, and H. Park. Gmar: gradient-driven multi-head attention rollout for vision transformer interpretability. In *2025 IEEE International Conference on Image Processing (ICIP)*, pages 582–587. IEEE, 2025.

[31] X. Ju, X. Liu, X. Wang, Y. Bian, Y. Shan, and Q. Xu. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In *European Conference on Computer Vision*, pages 150–168. Springer, 2024.

[32] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *2009 IEEE 12th international conference on computer vision*, pages 2106–2113. IEEE, 2009.

[33] G. T. A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ram'e, M. Rivière, L. Rouillard, T. Mesnard, G. Cideron, J.-B. Grill, S. Ramos, E. Yvinec, M. Casbon, E. Pot, I. Penchev, G. Liu, F. Visin, K. Kenealy, L. Beyer, X. Zhai, A. Tsitsulin, R. I. Busa-Fekete, A. Feng, N. Sachdeva, B. Coleman, Y. Gao, B. Mustafa, I. Barr, E. Parisotto, D. Tian, M. Eyal, C. Cherry, J.-T. Peter, D. Sinopalnikov, S. Bhupatiraju, R. Agarwal, M. Kazemi, D. Malkin, R. Kumar, D. Vilar, I. Brusilovsky, J. Luo, A. Steiner, A. Friesen, A. Sharma, A. Sharma, A. M. Gilady, A. Goedeckemeyer, A. Saade, A. Kolesnikov, A. Bendebury, A. Abdagic, A. Vadi, A. Gyorgy, A. S. Pinto, A. Das, A. Bapna, A. Miech, A. Yang, A. Paterson, A. Shenoy, A. Chakrabarti, B. Piot, B. Wu, B. Shahriari, B. Petrini, C. Chen, C. L. Lan, C. A. Choquette-Choo, C. Carey, C. Brick, D. Deutsch, D. Eisenbud, D. Cattle, D. Cheng, D. Paparas, D. S. Sreepathihalli, D. Reid, D. Tran, D. Zelle, E. Noland, E. Huizenga, E. Kharitonov, F. Liu, G. Amirkhanyan, G. Cameron, H. Hashemi, H. Klimczak-Pluci'nska,

H. Singh, H. Mehta, H. T. Lehri, H. Hazimeh, I. Ballantyne, I. Szpektor, I. Nardini, J. Pouget-Abadie, J. Chan, J. Stanton, J. M. Wieting, J. Lai, J. Orbay, J. Fernandez, J. Newlan, J. Ji, J. Singh, K. Black, K. Yu, K. Hui, K. Vodrahalli, K. Greff, L. Qiu, M. Valentine, M. Coelho, M. Ritter, M. Hoffman, M. Watson, M. Chaturvedi, M. Moynihan, M. Ma, N. Babar, N. Noy, N. Byrd, N. Roy, N. Momchev, N. Chauhan, O. Bunyan, P. Botarda, P. Caron, P. K. Rubenstein, P. Culliton, P. Schmid, P. G. Sessa, P. mei Xu, P. Stańczyk, P. D. Tafti, R. Shivanna, R. Wu, R. Pan, R. A. Rokni, R. Willoughby, R. Vallu, R. Mullins, S. Jerome, S. Smoot, S. Girgin, S. Iqbal, S. Reddy, S. Sheth, S. Põder, S. Bhatnagar, S. R. Panyam, S. Eiger, S. Zhang, T. Liu, T. Yacovone, T. Liechty, U. Kalra, U. Evci, V. Misra, V. Roseberry, V. Feinberg, V. Kolesnikov, W. Han, W. Kwon, X. Chen, Y. Chow, Y. Zhu, Z. Wei, Z. Egyed, V. Cotruta, M. Giang, P. Kirk, A. Rao, J. Lo, E. Moreira, L. G. Martins, O. Sanseviero, L. Gonzalez, Z. Gleicher, T. Warkentin, V. S. Mirrokni, E. Senter, E. Collins, J. Barral, Z. Ghahramani, R. Hadsell, Y. Matias, D. Sculley, S. Petrov, N. Fiedel, N. Shazeer, O. Vinyals, J. Dean, D. Hassabis, K. Kavukcuoglu, C. Farabet, E. Buchatskaya, J.-B. Alayrac, R. Anil, D. Lepikhin, S. Borgeaud, O. Bachem, A. Joulin, A. Andreev, C. Hardin, R. Dadashi, and L. Hussenot. Gemma 3 technical report. *ArXiv*, abs/2503.19786, 2025.

[34] J. Lee, F. Chen, S. Dua, D. Cer, M. Shanbhogue, I. Naim, G. H. Ábrego, Z. Li, K. Chen, H. S. Vera, et al. Gemini embedding: Generalizable embeddings from gemini. *arXiv preprint arXiv:2503.07891*, 2025.

[35] S. Leem and H. Seo. Attention guided cam: visual explanations of vision transformer guided by self-attention. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 2956–2964, 2024.

[36] T. Li, Z. Wen, L. Song, J. Liu, Z. Jing, and T. S. Lee. From local cues to global percepts: Emergent gestalt organization in self-supervised vision models. *arXiv preprint arXiv:2506.00718*, 2025.

[37] W. Li, Z. Lin, K. Zhou, L. Qi, Y. Wang, and J. Jia. Mat: Mask-aware transformer for large hole image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10758–10768, 2022.

[38] Y. Li, Y. Du, K. Zhou, J. Wang, X. Zhao, and J.-R. Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 292–305, 2023.

[39] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[40] N. Liu, N. Zhang, K. Wan, L. Shao, and J. Han. Visual saliency transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4702–4712. IEEE, 2021.

[41] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[42] I. Misra, C. Lawrence Zitnick, M. Mitchell, and R. Girshick. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2930–2939, 2016.

[43] S. Murlidaran and M. P. Eckstein. Evolution of accuracy and visual-cognitive errors in a decade of vision-language ai models. *arXiv preprint arXiv:2607.09654*, 2026.

[44] S. E. Palmer. *Vision science: Photons to phenomenology*. MIT press, 1999.

[45] W. Peebles and S. Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[46] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. *Pattern recognition*, 106:107404, 2020.

[47] M. T. Ribeiro, S. Singh, and C. Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[48] A. Rohrbach, L. A. Hendricks, K. Burns, T. Darrell, and K. Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, 2018.

[49] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[50] L. Rout, Y. Chen, N. Ruiz, C. Caramanis, S. Shakkottai, and W.-S. Chu. Semantic image inversion and editing using rectified stochastic differential equations. *arXiv preprint arXiv:2410.10792*, 2024.

[51] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[52] A. Singh, A. Fry, A. Perelman, A. Tart, A. Ganesh, A. El-Kishky, A. McLaughlin, A. Low, A. Ostrow, A. Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

[53] N. Spanos, M. Lymperaiou, G. Filandrianos, K. Thomas, A. Voulodimos, and G. Stamou. V-cece: Visual counterfactual explanations via conceptual edits. *arXiv preprint arXiv:2509.16567*, 2025.

[54] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, and V. Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2149–2159, 2022.

[55] M. R. Taesiri, G. Nguyen, S. Habchi, C.-P. Bezemer, and A. Nguyen. Imagenet-hard: The hardest images remaining from a study of the power of zoom and spatial biases in image classification. *Advances in Neural Information Processing Systems*, 36:35878–35953, 2023.

[56] B. W. Tatler. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of vision*, 7(14):4–4, 2007.

[57] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[58] S. Tong, Z. Liu, Y. Zhai, Y. Ma, Y. LeCun, and S. Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9568–9578, 2024.

[59] S. Ullman, L. Assif, E. Fetaya, and D. Harari. Atoms of recognition in human and computer vision. *Proceedings of the National Academy of Sciences*, 113(10):2744–2749, 2016.

[60] B. Wallace, M. Dang, R. Rafailov, L. Zhou, A. Lou, S. Purushwalkam, S. Ermon, C. Xiong, S. Joty, and N. Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, 2024.

[61] A. N. Wang, C. Hoang, Y. Xiong, Y. LeCun, and M. Ren. Poodle: Pooled and dense self-supervised learning from naturalistic videos. *arXiv preprint arXiv:2408.11208*, 2024.

[62] Z. Wen, T. Li, Z. Jing, and T. S. Lee. Does resistance to style-transfer equal global shape bias? measuring network sensitivity to global shape configuration. *arXiv preprint arXiv:2310.07555*, 2023.

[63] E. Xie, J. Chen, J. Chen, H. Cai, H. Tang, Y. Lin, Z. Zhang, M. Li, L. Zhu, Y. Lu, et al. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv preprint arXiv:2410.10629*, 2024.

[64] J. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, and Q. Zhao. Predicting human gaze beyond pixels. *Journal of vision*, 14(1):28–28, 2014.

[65] M. Yuksekgonul, F. Bianchi, P. Kalluri, D. Jurafsky, and J. Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022.

[66] D. Zhang, J. Li, Z. Zeng, and F. Wang. Jasper and stella: distillation of sota embedding models. *arXiv preprint arXiv:2412.19048*, 2024.

[67] Y. Zhang, M. Li, D. Long, X. Zhang, H. Lin, B. Yang, P. Xie, A. Yang, D. Liu, J. Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

[68] J. Zhu, W. Wang, Z. Chen, Z. Liu, S. Ye, L. Gu, H. Tian, Y. Duan, W. Su, J. Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

[69] J. Zhuang, Y. Zeng, W. Liu, C. Yuan, and K. Chen. A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. In *European Conference on Computer Vision*, pages 195–211. Springer, 2024.

A Technical Details of Counterfactual Semantic Saliency

This section shows the prompts and hyperparameters employed in the Counterfactual Semantic Saliency (CSS) pipeline.

For the object name generation stage, we utilized GPT-4o[1]. For each factual scene, we sampled two independent responses using the prompt detailed below to maximize object coverage. Empirical testing indicated that sampling more than twice yielded negligible marginal gains in discovering distinct, new objects. The sampling temperature was set to 1.0, with all other hyperparameters at their default values.

Crucially, the objective of this approach is not the exhaustive enumeration of every entity in the scene. Rather, it is to generate a controlled counterfactual environment for robust human and model evaluation. Consequently, guaranteeing absolute recall of all objects is not strictly necessary for our methodology.

Prompts for Object List Generation

List the names of all objects in the scene, do NOT mention the objects that would be considered as background (like sky, sea, ground), each separated by a comma. The name of the object should be short. The format MUST looks like: 'red car, blue car, man in black'. If you cannot identify any object, please write 'None'.

Prompts for Object List Merging

I have {len(caption_list)} lists of object names from the same image. Please merge them into one list, removing duplicates.
Objects that refer to the same thing should be considered duplicates even if worded slightly differently (e.g., "man in black" and "person in black" are duplicates). Keep the most descriptive version when there are duplicates. If the object name contains two entries (e.g., "man with umbrella"), split them into two separate objects ("man with umbrella" and "umbrella").
{list_lines}
Please return the merged list as a comma-separated string in the same format as the input. For example: 'red car, blue car, man in black'. If all lists contain "None" or no valid objects, return "None". You should ONLY return the list of object names, NO other text.

Object removal was executed using the Nano Banana 2 model via the Gemini API. To guide the generative inpainting process and ensure the semantic integrity of the underlying scene, we overlaid the factual scene with a transparent red mask targeting the specified object. Such masks are obtained through the SOTA text-prompted segmentation model SAM3 [6], see the main content for how the masks are processed before being applied to the Object Removal Stage. The segmentation process is computed by using a single GeForce RTX 4090 GPU.

Prompts for Object Removal

Remove the object covered with red masks; ONLY remove the object itself.

During the VLM evaluation using the Counterfactual Semantic Saliency, we applied the prompt detailed below to all evaluated models for every image. To account for generation variability, we sampled five independent descriptions per image. An ablation study comparing these results against deterministic sampling is presented in the subsequent section.

For all models, the maximum generation length was set to 16,384 tokens to ensure the capture of complete descriptions. For such reasoning models, only the final generated scene description was extracted for evaluation. All other sampling parameters were maintained at their default settings, whether the models were accessed via Hugging Face or API call. For open-source model evaluations, all are executed on a single NVIDIA RTX PRO 6000 GPU.

Prompts for Scene Description

Make your best guess of what might be happening in this scene in one sentence. Avoid mentioning objects that do not aid in understanding the context of the scene.

B Data Quality Validation Protocols

To ensure the high fidelity of our counterfactual scene generation, we recruited 14 human annotators via Prolific to verify the successful removal of target objects. Annotators were presented with side-by-side image pairs: the original factual scene is covered with a red mask highlighting the target object, and the generated counterfactual scene where the object was intended to be ablated.

To maintain annotation quality, the generated images were divided into 7 distinct batches, capping each batch at fewer than 300 trials. Two independent annotators were assigned to evaluate each batch. Their specific instructions for identifying successful object ablation are detailed below:

Instructions for Data Quality Validation

In this experiment, each time you will be given a pair of images. The image on the right is an image with a red mask on a particular object. The image on the left is the image where that object is removed.
Your task is to identify whether the object with a red mask is removed in the left image. The answer is always "Yes" or "No":
When the object with red masks is successfully removed, select “Yes”.
When the object with red masks is not successfully removed, select “No”.

Prior to the main task, annotators were provided with standardized examples of successful (“Yes”) and failed (“No”) object removals as illustrated in Fig.7. During the experiment, stimuli were presented using this exact standardized layout.

Upon collection of the crowdsourced annotations, any counterfactual image flagged with a “No” by either of the two assigned annotators underwent a secondary manual review by the researchers. Images definitively exhibiting failed ablations were discarded. Furthermore, if the discarding of failed generations caused a factual scene to fall below a minimum threshold of three valid counterfactual variants, the entire factual scene and its remaining variants were removed from the dataset. Following this filtering protocol, the finalized dataset comprises 307 unique factual scenes and 1,306 high-fidelity counterfactual variants.

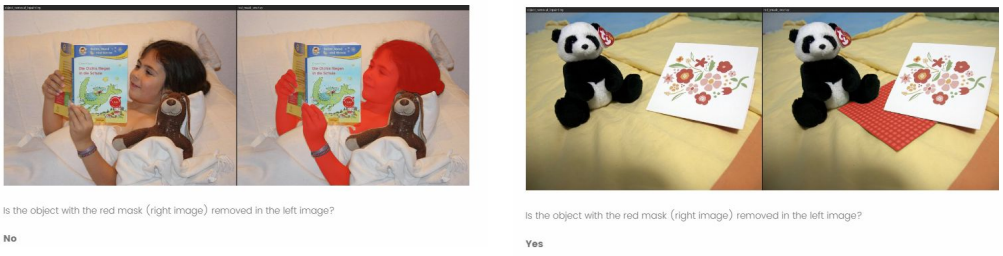


Figure 7: Examples provided for annotators in data quality validation process.

C Online Psychophysics Details

This section contains the explicit instructions for our large-scale online data collection. A comprehensive overview of the core psychophysical experimental design is provided in the main text.

Participants were recruited via the Prolific crowdsourcing platform. To ensure high-quality textual descriptions, eligibility was strictly restricted to native or highly fluent English speakers who maintained a historical task approval rate of $\geq 98\%$. To ensure strict methodological parity between the

human baseline and the models, the primary task instruction presented to human participants, *"Make your best guess of what might be happening in this scene in one sentence. Avoid mentioning objects that do not aid in understanding the context of the scene."*, was identical to the text prompt utilized for the evaluated Vision-Language Models.

Instructions for Online Psychophysics

Your task will be to look at the image and carefully provide a description.
Make your best guess of what might be happening in this scene in one sentence. Avoid mentioning objects that do not aid in understanding the context of the scene.
We expect the following for the descriptions:
1. The description should explain what is happening in the scene.
2. The descriptions need to be concise.
The first three images have been provided with descriptions as an example. Please carefully review the examples, as they will give you an idea of the kind of images you will see in the survey and the kind of descriptions we expect.
You need to satisfy these requirements to participate:
1. You MUST be a Native English speaker. This means that you were raised speaking English
2. You MUST carefully look at the example shown and provide descriptions as suggested.
3. You MUST thoroughly review each image and provide a meaningful and grammatically correct description.
4. Please ensure to open this link on a laptop or Desktop.

To standardize the verbosity of human responses, participants were required to review examples of acceptable textual descriptions before beginning the main experimental trials (Fig. 8). Importantly, the visual stimuli utilized for these calibration examples were strictly disjoint from the main dataset used for the formal human study and the VLM evaluations. Subjects can have both factual and counterfactual scenes in their psychophysics, while they are not informed that some of the scenes are generated object-ablated scenes.

Following data collection, participants identified as potential automated bots or those exhibiting consistently low-effort responses were completely excluded from the dataset. Furthermore, responses were discarded if their mean semantic similarity to other human responses for the same image was below 0.4, as measured by sentence embedding [66].

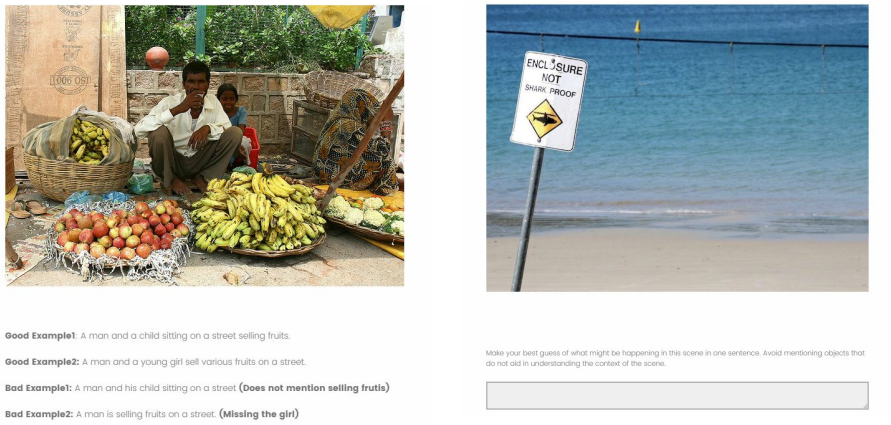


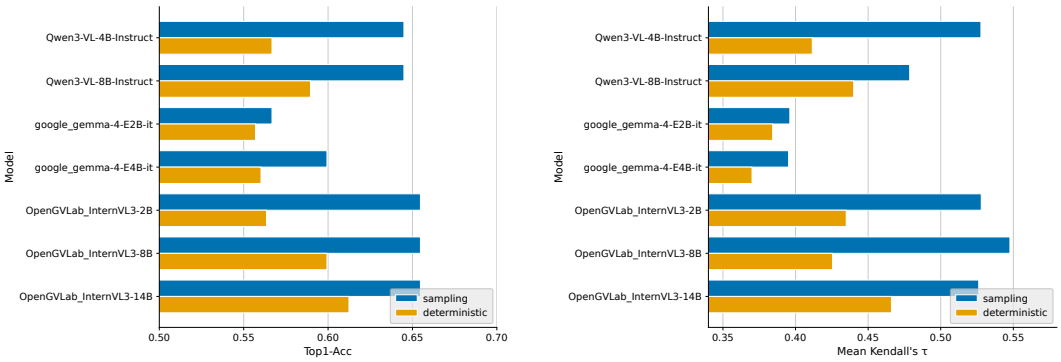
Figure 8: **Examples of Online Psychophysics**: Left: One of the examples provided to subjects as calibrations. Right: One of the trails within the experiment.

D Ablations: Deterministic vs. Stochastic Decoding

As detailed in the main text, our primary evaluation employs stochastic sampling during VLM generation to capture the variance and distribution inherent in human scene perception. In this section,

we present an ablation study investigating the impact of restricting models to a single deterministic output. Specifically, we re-evaluate the seven open-source models featured in our primary results (e.g., Qwen3-VL, Gemma4, and InternVL3) by generating their textual descriptions using greedy decoding. We then quantify their semantic alignment with the human cognitive baseline using Top-1 Accuracy and Kendall's τ.

As illustrated in Fig.9, models evaluated under deterministic decoding exhibit a significant performance drop compared to their sampling counterparts. This drop in human-AI alignment indicates that introducing variability via stochastic sampling is necessary for models to accurately approximate the true distribution of human responses.



(a) Top-1 Accuracy at predicting the critical object.

(b) Mean Kendall's τ between models and humans.

Figure 9: Comparison between deterministic output and stochastic sampling. *Sampling* results are the same results in the main paper.

E Robustness across Inpainting Methods

Because CSS relies on generative models to construct counterfactual scenes, the inpainting method itself could be a confound. We consider two such confounds. First, the inpainting model reconstructs the masked region according to its own prior, so the observed semantic shift might partially reflect properties of the inpainting model rather than the removal of the target object. Second, our analysis assumes that object removal is the only meaningful difference between a factual scene and its counterfactual variant. A generative model such as Nano Banana 2 may also alter content outside the target mask, and such changes could correlate with object size. In this section, we address each concern in turn and show that neither accounts for the model–human gap or the attribute sensitivity reported in the main text.

E.1 Replacing the Inpainting Model

To test whether our findings depend on the prior of a specific inpainting model, we regenerated all counterfactual stimuli using the Mask-Aware Transformer (MAT) [37], an inpainting model based on GAN whose generative prior differs fundamentally from that of the diffusion-based Nano Banana 2. Although MAT may introduce visible artifacts in some images, its distinct prior makes it a useful test of whether the scene comprehension gap and the attribute sensitivity are specific to a particular inpainting model. Using the same segmentation masks and preprocessing pipeline as in the main experiments, we removed the same target objects and obtained a new set of 1,306 counterfactual scenes. We then repeated the full evaluation of all 19 VLMs, including the Top-1 Accuracy analysis (Fig. 4a) and the four attribute sensitivity analyses (Fig. 5).

As shown in Fig. 10, the systematic gap between human–human consistency and model Top-1 Accuracy persists and remains statistically significant under MAT inpainting. Likewise, the divergences in all four attribute sensitivities are preserved (Fig. 11): models continue to rely more on object size, object in the center, and Low-level Saliency, while being less sensitive to whether the object is a person. These results indicate that the alignment gaps are not artifacts of the Nano Banana 2 prior.

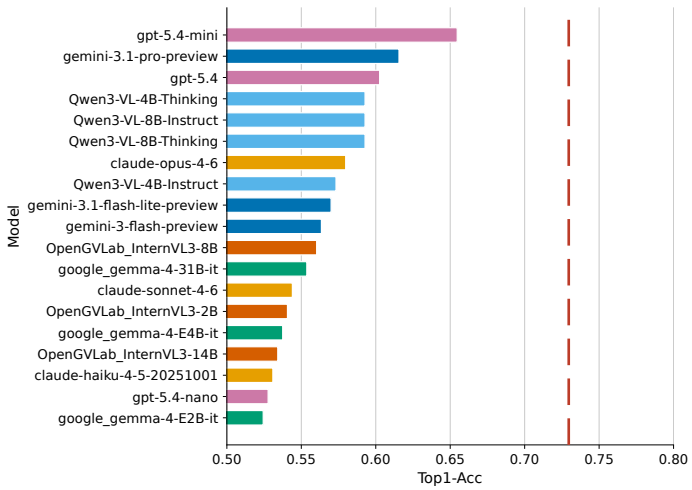


Figure 10: Top-1 Accuracy at predicting the critical object when counterfactual scenes are generated with MAT [37]. The dashed line denotes human–human consistency. All gaps are significant (one-tailed bootstrap test, $p < .01$).

E.2 Changes Outside the Object Mask

Generative editing models may alter image content beyond the target region. For example, in Fig. 2, removing the TV also removes the lighting effect of the pink LED on the surrounding wall. Although our human validation (Appendix B) ensures that the target objects are successfully removed, it does not rule out the possibility that such out-of-mask changes correlate with object size and thereby contribute to the observed over-reliance to size.

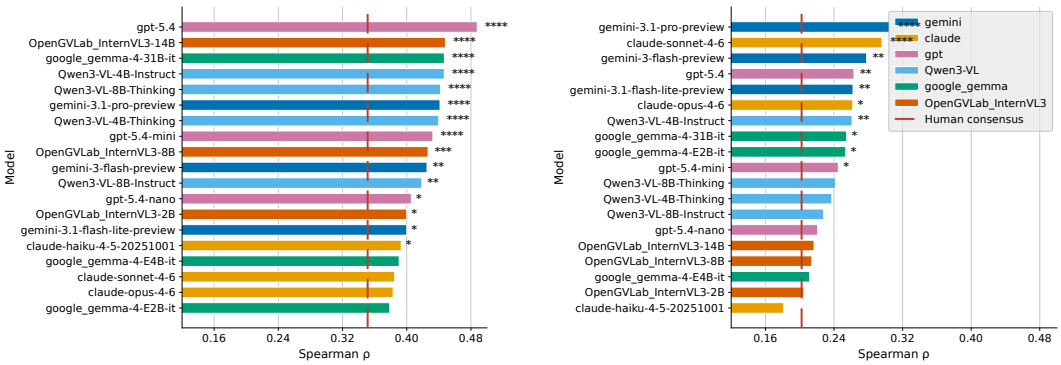
Measuring out-of-mask changes. To quantify changes outside the target region without confounding them with mask size, we compute a modified Peak Signal-to-Noise Ratio (PSNR) that excludes the masked region. Let I and I' denote the factual and counterfactual images, and let $\tilde{M}$ denote the target mask dilated with a 24×24 kernel, identical to the dilation applied when preparing the inputs to the inpainting model. The mean squared error is computed only over the pixels outside the dilated mask, and we refer to the resulting metric as $\mathrm{PSNR}_{\mathrm{out}}$. Across all 1,306 counterfactual scenes generated by Nano Banana 2, PSNRout is only weakly correlated with mask size (Pearson r=-0.166), suggesting that the size of the removed object has a minor influence on changes to the rest of the scene.

Out-of-mask changes do not explain the Sensitivity to Size. If the Sensitivity to Size reported in Fig. 5a were driven by out-of-mask changes, these changes should affect models and humans differently. In that case, replacing object size with $\mathrm{PSNR}_{\mathrm{out}}$ in the sensitivity analysis should reproduce a similarly systematic and significant model–human gap. We therefore repeated the analysis of Fig. 5, computing the Spearman correlation between $\mathrm{PSNR}_{\mathrm{out}}$ and CSS for each model and for humans (human $\rho = 0.210$).

As shown in Fig. 12, the pattern differs markedly from that of the Sensitivity to Size. With the exception of gpt-5.4-nano, no model differs significantly from humans in its sensitivity to out-of-mask changes. Moreover, gpt-5.4-nano shows a significantly *weaker* correlation than humans in this analysis, whereas it shows a significantly *stronger* Sensitivity to Size than humans in Fig. 5a. Thus, out-of-mask changes affect models and humans similarly and cannot account for the divergence of sensitivity to size observed in the main text.

F Robustness across Sentence Embedding Models

Here we show that the divergence between model and human through CSS is robust to the change of sentence embedding model. We replaced the embedding model used in the main experiments, Jasper (`jasper_en_vision_language_v1`) [66], with two alternatives: the open-source Qwen3-



(a) Sensitivity to Size

(b) Sensitivity to Center

(c) Sensitivity to Low-level Saliency

(d) Sensitivity to Person-Non-person

Figure 11: Attribute Sensitivity of VLMs when counterfactual scenes are generated with MAT [37].

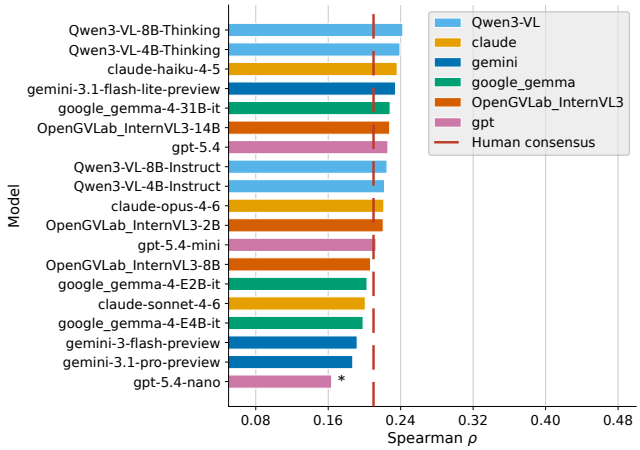


Figure 12: Spearman correlation between the modified PSNR outside the object mask (PSNR_{out}) and CSS for each evaluated model.

Embedding-0.6B [67] and the closed-source gemini-embedding-001 [34]. We examined the robustness of our results at two levels: the CSS scores themselves (Sec. F.1) and the model–human comparisons (Sec. F.2).

F.1 Consistency of CSS Scores

For each evaluated VLM and for the human baseline, we recomputed the CSS score of every counterfactual scene (1,306 in total) using each of the three embedding models. We then measured

the Spearman rank correlation between the CSS scores obtained with each pair of embedding models to check if changing the embedding model has a non-negligible impact on the CSS score. As shown in Table 3, the rank order of CSS scores changes little across embedding models. For all VLMs and for humans, the Spearman correlations between embedding models are consistently close to 0.95, indicating that CSS itself is robust to the choice of sentence embedding model.

Table 3: Spearman rank correlation of CSS scores computed with different sentence embedding models across all 1,306 counterfactual scenes. J: Jasper (used in the main experiments); Q: Qwen3-Embedding-0.6B; G: gemini-embedding-001.

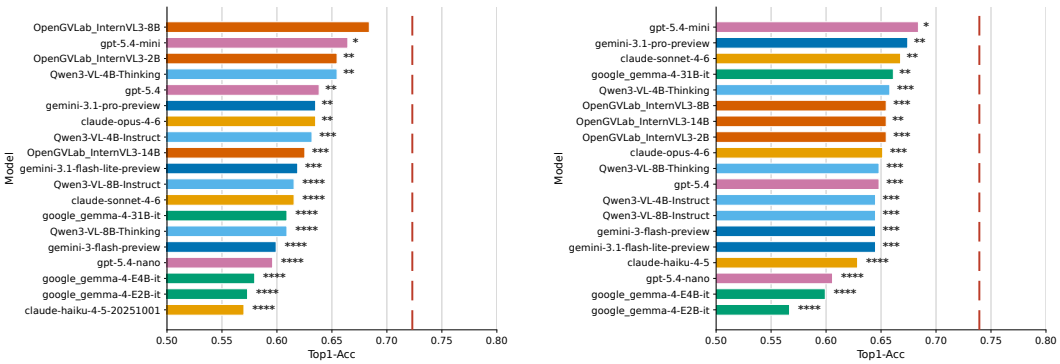
Source	J vs. Q	J vs. G	Q vs. G
Human	0.944	0.945	0.933
claude-haiku-4-5	0.965	0.965	0.956
claude-opus-4-6	0.957	0.957	0.949
claude-sonnet-4-6	0.963	0.954	0.942
gemini-3-flash-preview	0.958	0.962	0.956
gemini-3.1-flash-lite-preview	0.961	0.963	0.958
gemini-3.1-pro-preview	0.966	0.962	0.957
gemma-4-31B-it	0.968	0.975	0.964
gemma-4-E2B-it	0.968	0.975	0.965
gemma-4-E4B-it	0.958	0.969	0.951
gpt-5.4	0.966	0.973	0.963
gpt-5.4-mini	0.963	0.970	0.961
gpt-5.4-nano	0.952	0.958	0.953
InternVL3-2B	0.938	0.938	0.926
InternVL3-8B	0.946	0.958	0.940
InternVL3-14B	0.946	0.958	0.938
Qwen3-VL-4B-Instruct	0.965	0.968	0.963
Qwen3-VL-4B-Thinking	0.965	0.968	0.960
Qwen3-VL-8B-Instruct	0.968	0.970	0.963
Qwen3-VL-8B-Thinking	0.969	0.972	0.964

F.2 Stability of the Model–Human Gap

Beyond the consistency of the scores themselves, we examined whether our main conclusions hold with different choices of embedding models. Using Qwen3-Embedding-0.6B and gemini-embedding-001, we repeated the Top-1 Accuracy analysis (Fig. 4a) and the four attribute sensitivity analyses (Fig. 5), with human–human consistency and the human sensitivity baselines recomputed in the corresponding embedding space.

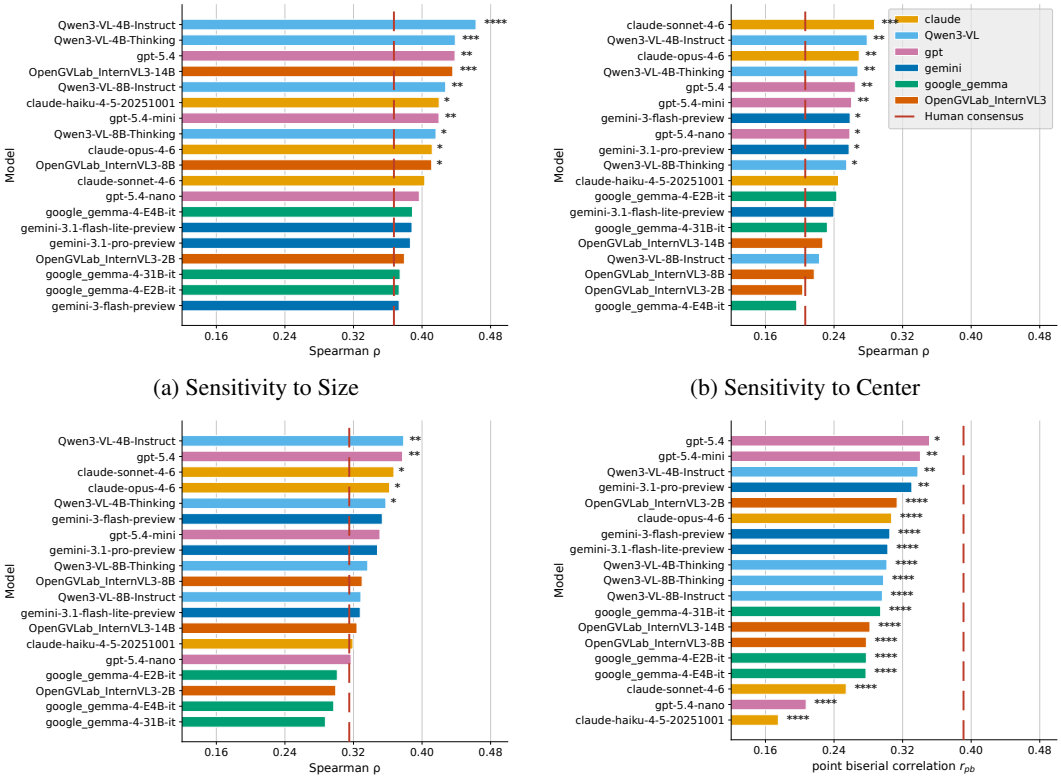
Accuracy gap. As shown in Fig. 13, the systematic gap between human–human consistency and model Top-1 Accuracy persists under both alternative embedding models, consistent with the results obtained using Jasper.

Attribute Sensitivity gap. Fig. 14 and Fig. 15 report the four attribute sensitivities under Qwen3-Embedding-0.6B and gemini-embedding-001, respectively. Together with the high consistency of CSS scores across embedding models, these results show that our findings are not artifacts of a particular sentence embedding model. In both cases, the divergences observed in the main text are mostly preserved: compared with humans, models are more sensitive to Size, Center, and Low-level Saliency, and insensitive to whether the object is a person or not, except for the sensitivity to low-level saliency measured under Qwen3-Embedding-0.6B, where there remain fewer models that are significantly different from the human baseline.



(a) Qwen3-Embedding-0.6B (b) gemini-embedding-001

Figure 13: Top-1 Accuracy at predicting the critical object when CSS is computed with different sentence embedding models. The dashed line denotes human–human consistency computed with the same embedding model.

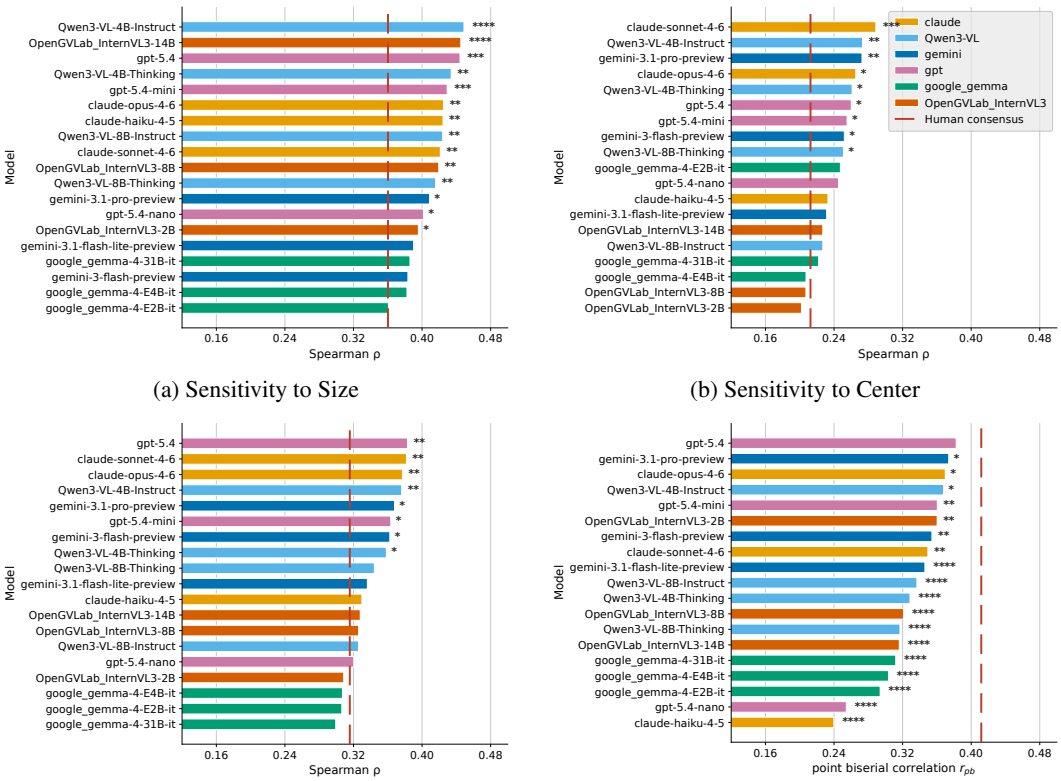


(c) Sensitivity to Low-level Saliency (d) Sensitivity to Person-Non-person

Figure 14: Attribute Sensitivity of VLMs when CSS is computed with Qwen3-Embedding-0.6B.

G Variability in Human Agreement Does Not Account for the Model–Human Divergence

Not every scene necessarily has a robust human consensus. For some images, observers may disagree about which object is most critical to the scene, so the ground truth itself is uncertain. One possibility is that the gap between humans and models is due to the variability among human observers rather than a divergence in scene comprehension. Here, we test whether the gap persists after controlling for the certainty of human agreement.



(a) Sensitivity to Size

(b) Sensitivity to Center

(c) Sensitivity to Low-level Saliency

(d) Sensitivity to Person-non-person

Figure 15: Attribute Sensitivity of VLMs when CSS is computed with gemini-embedding-001.

Quantifying human certainty. Because our psychophysics experiment follows a between-subjects design, each counterfactual variant of a scene was described by a different group of participants. Individual differences therefore cannot be measured directly. Instead, we estimate the certainty of human agreement for each scene as follows. For each condition of a scene (the factual image and each of its counterfactual variants), we randomly select a single participant's description from the 10 participants assigned to that condition. Using these single-participant descriptions, we compute the CSS of each object and identify the critical object as the one with the highest CSS. We repeat this sampling procedure and define the human certainty score of the scene as the probability that the critical object obtained from single participants matches the critical object determined from all participants' responses. A higher score indicates that the identification of the critical object is consistent across individual observers, whereas a lower score indicates that the human consensus is less stable.

Results. We split the 307 factual scenes into four quartile bins according to their human certainty scores (77 scenes per bin, except Q3, which contains 76). Within each bin, we report the Top-1 Accuracy averaged across all 19 evaluated VLMs alongside the human–human consistency, computed using the same procedure described in Sec. 4.1. Error bars denote the standard deviation of the bootstrap distribution, and statistical significance is assessed with a one-tailed bootstrap test of whether human–human consistency exceeds the mean model accuracy.

As shown in Table 4, models and humans show no significant difference only in the lowest-certainty bin (Q1). However, in all three remaining bins, human–human consistency significantly exceeds the mean model accuracy. Notably, the gap remains significant even in Q2, and it persists through the highest-certainty bins (Q3 and Q4), indicating that the model–human divergence cannot be explained simply by the variability of human agreement.

Table 4: Model–human gap in critical object identification across levels of human certainty. Scenes are split into quartiles based on the certainty of human agreement. Mean Model Top-1 Acc is averaged across all 19 evaluated VLMs. Values are reported as mean $\pm$ bootstrap standard deviation; p-values are obtained from one-tailed bootstrap tests, indicating whether the human–model gap is significant.

Certainty Bin	Certainty Range	Mean Model Top-1 Acc	Human–Human Consistency	Gap p
Q1	$[0.226, 0.462]$	0.338 ± 0.029	0.286 ± 0.052	0.807
Q2	$(0.462, 0.635]$	0.506 ± 0.037	0.649 ± 0.055	0.008
Q3	$(0.635, 0.850]$	0.735 ± 0.026	0.987 ± 0.013	< 0.001
Q4	$(0.850, 1.000]$	0.918 ± 0.018	1.000 ± 0.000	< 0.001

H Extended Results: Model-Human Ranking Correlations

Due to space constraints, the main text reports only the aggregated mean Kendall's τ coefficients for the object importance rankings between models and human observers across the dataset. However, because Kendall's τ is computed independently for each factual scene, examining the full distribution of these coefficients provides a much more granular perspective on model–human alignment. In this section, we present the complete distributions of Kendall's τ for each evaluated Vision-Language Model (VLM) when compared against the human baseline (Fig. 16). The human–human consistency is also illustrated at the end of the figure (computed in the same way as described in Sec. 4.1).

I Extended Qualitative Comparisons: CSS vs. White-Box Saliency

In this section, we provide extended qualitative visualizations comparing white-box interpretability methods against our causal Counterfactual Semantic Saliency (CSS) framework (Fig.17). The activation values within the CSS maps are globally normalized across all factual scenes in our dataset to standardize the scale of the induced semantic shift.

J Broader Impacts

Understanding the divergence in scene perception is critical for AI safety. We reveal that modern VLMs overly rely on spatial dominance rather than top-down semantic importance (such as the presence of humans). Identifying and mitigating this unnatural sensitivity is crucial before deploying VLMs in high-stakes environments, such as autonomous driving or medical imaging, where ignoring a small but semantically critical object (e.g., a pedestrian) in favor of a large background element could have catastrophic consequences. Conversely, exposing these specific mechanistic vulnerabilities could be exploited by malicious actors. By demonstrating the severity of VLM's reliance to Size and Center, adversarial attacks could be designed using physically large, irrelevant distractors to intentionally blind AI security systems to actual targets.

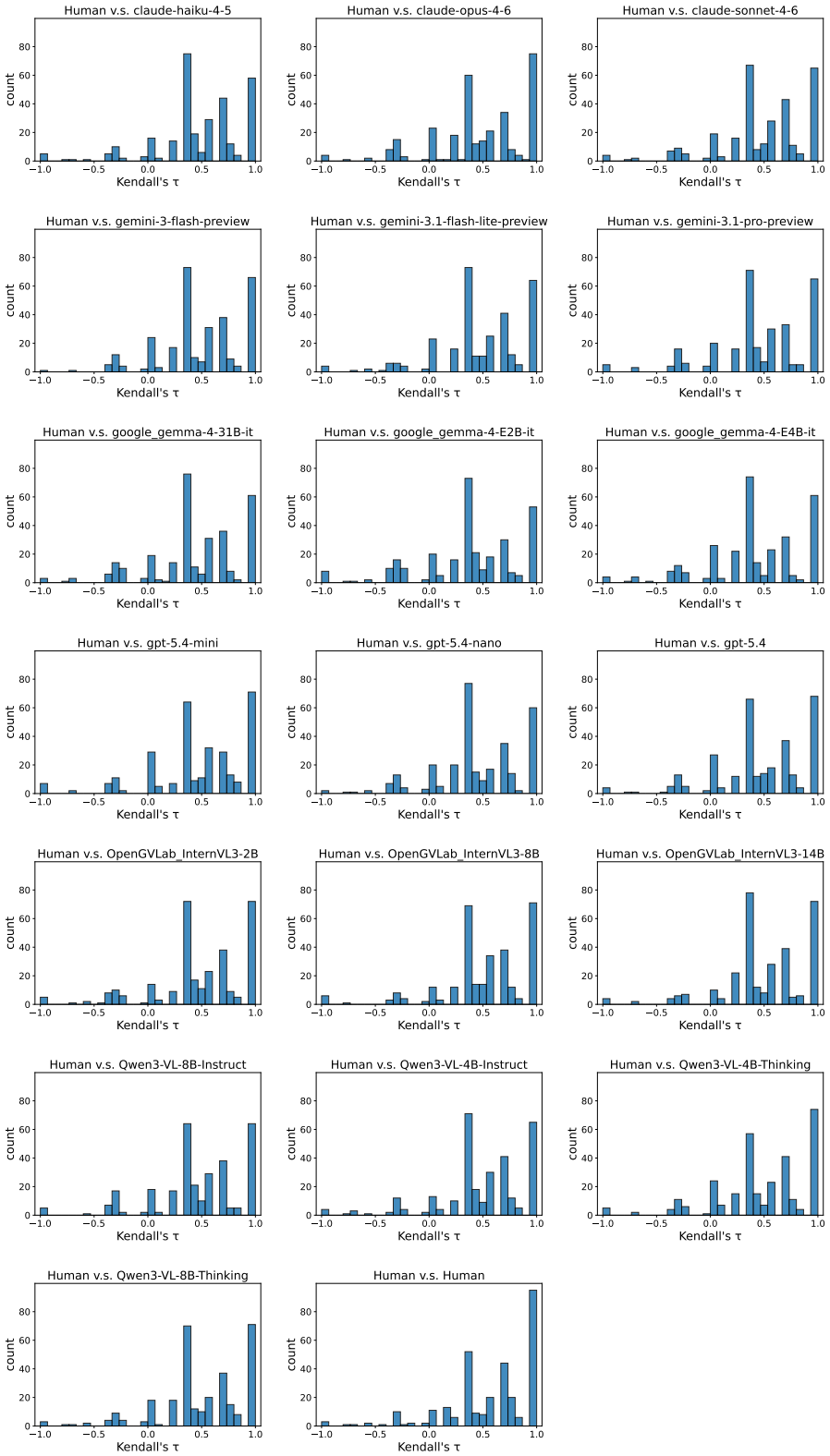


Figure 16: Distribution of Kendall's τ for each model comparing against human, and human-human consistency

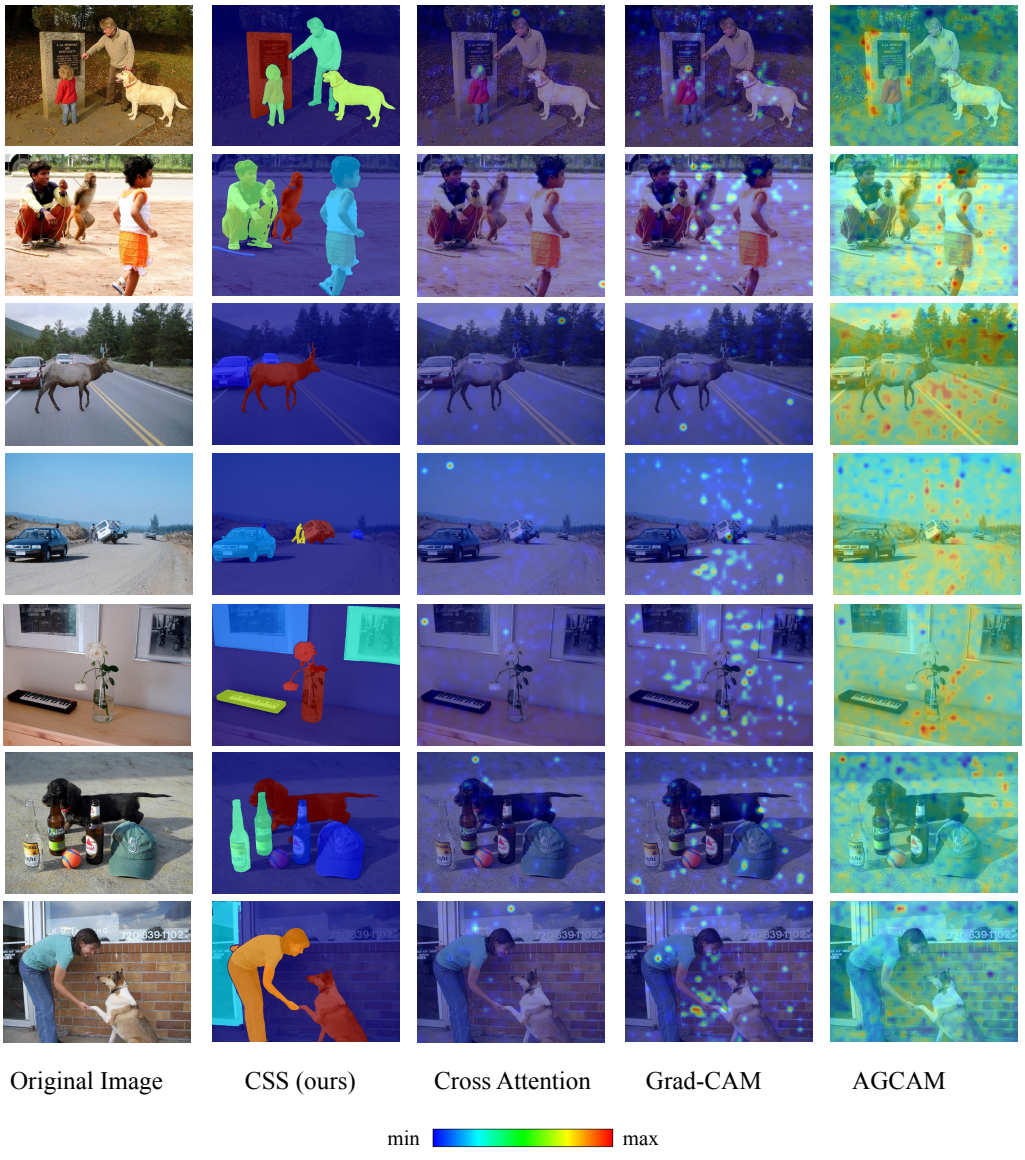


Figure 17: Comparisons between CSS Maps and attention-based white-box maps

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, all the claims made in the abstract and introduction have been verified through experiment and analysis in the Results section

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitation is discussed in the Discussion section.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper does not contain the theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Yes, data and scripts needed for reproducing the results in the paper are included in the anonymized GitHub URL.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Yes, the data and scripts to reproduce the results in the paper are put in the anonymized GitHub URL in the abstract.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, the dataset selection is explained in the main paper, Method section. All the details related to the model are included in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Yes, as mentioned in the Results Section, statistical significance is provided by one-tailed bootstrap tests or permutation tests.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes, the computer resources used are specified in the first section of the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, the research follows the NeurIPS Code of Ethics

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Due to the limitation of pages in the main content, we put the broader impact section in the appendix.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not contain any new trained model, and the dataset is based on a publicly released dataset. No potentially sensitive content exists within the dataset.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes, all third-party code, models, and datasets are open-source. We provide citations to the original creators.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Yes, we introduce a new dataset for evaluating the human-ai scene perception gap. The dataset collection process is discussed in the paper, Method section. And the link to download the dataset can be accessed with an anonymized GitHub URL in the abstract.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Yes, full instructions are included in the appendix.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: Yes. All human subject experiments were approved by our institution's IRB.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Non-standard usage of LLMs has been explicitly explained in the Method section and the Result Section.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.