The Perceptual Observatory Characterizing Robustness and Grounding in MLLMs

Tejas Anvekar* Fenil Bardoliya* Pavan K. Turaga Chitta Baral Vivek Gupta
Arizona State University
{tanvekar, fbardoli, pturaga, chitta, vgupt140}@asu.edu
https://coral-lab-asu.github.io/PerceptualObservatory/

Abstract

Recent advances in multimodal large language models (MLLMs) have yielded increasingly powerful models, yet their perceptual capacities remain poorly characterized. In practice, most model families scale language component while reusing nearly identical vision encoders (e.g., Qwen2.5-VL 3B/7B/72B), which raises pivotal concerns about whether progress reflects genuine visual grounding or reliance on internet-scale textual world knowledge. Existing evaluation methods emphasize end-task accuracy, overlooking robustness, attribution fidelity, and reasoning under controlled perturbations. We present The Perceptual Observatory , a framework that characterizes MLLMs across verticals like: (i) simple vision tasks, such as face matching and text-in-vision comprehension capabilities; (ii) local-to-global understanding, encompassing image matching, grid pointing game, and attribute localization, which tests general visual grounding. Each vertical is instantiated with ground-truth datasets of faces and words, systematically perturbed through pixel-based augmentations and diffusion-based stylized illusions. The Perceptual Observatory moves beyond leaderboard accuracy to yield insights into how MLLMs preserve perceptual grounding and relational structure under perturbations, providing a principled foundation for analyzing strengths and weaknesses of current and future models.

1. Introduction

Multimodal Large Language Models (MLLMs) are ubiquitous for tasks such as captioning, VQA, OCR-centric reasoning, document understanding, accessibility, robotics, and multi-image dialogue [1, 3, 7, 9, 29, 31]. Public leaderboards (e.g., MMBench, MM-MU/Pro; TextVQA; VizWiz; SEEDBench; POPE; MATHVista) mostly report end-task accuracy [16, 26–28, 30, 32, 33, 41, 53]. However, outstanding benchmark performance does not guarantee robust *perception* – defined as the fundamental ability to faithfully

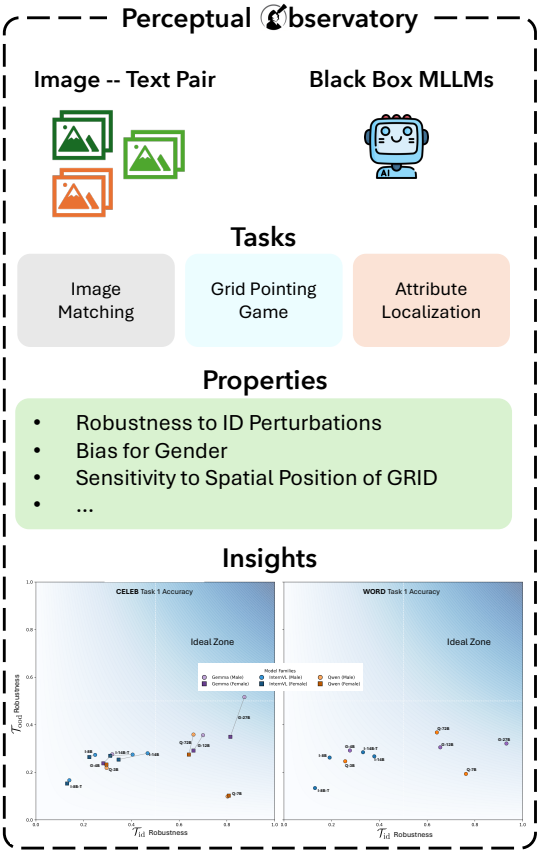

Figure 1. Overview of The Perceptual Observatory and how it solicits understanding of opaque MLLMs perceptual understanding by measuring properties motivated by human visual perception and robustness against multiple axes. We illustrate the framework for properties revealing true perceptual understanding of MLLMs

*contributed equally

understand and interpret visual details, maintain object identity, and spatially ground independent of linguistic reasoning. Without this, models can exploit textual priors, miss identity under perturbations, or fail to localize evidence.

Modern MLLMs scale the *language* side while leaving vision encoders frozen or lightly adapted via compact bridges (Qwen2.5-VL-family, Gemma3-family, Q-Former, Perceiver resampler, MLP/linear projectors) [3, 31, 42, 43]. This raises the question of whether the gains are due to better *visual* or better *textual* capabilities? Decades of vision research warns that models can rely on shortcuts; language priors in VQA or texture bias in CNNs masks poor perceptual grounding [2, 13, 15]. Furthermore, while web-scale pretraining increasingly obscures the boundary between In-distribution (ID) and out-of-distribution (OOD) data, foundational robustness studies demonstrate that accuracy can precipitously decline under even modest corruptions or distribution shifts [5, 19, 22]

Based on human-cognitive behaviour, where perception remains robust across stylistic variations and environmental noise [12, 25, 48, 49], we probe the depth of machine *seeing* against these biological standards. We then ask: (Q1) Do MLLMs *preserve identity* under content-preserving ID corruptions and under OOD *stylized images?* (Q2) Are predictions *positional-invariant* when the same content moves in a grid? (Q3) Do models *ground* attributes where they belong, and does giving hints improve transfer? (Q4) Does *scaling* primarily on the language side yield monotonic perceptual gains when the vision encoder is fixed? (Q5) Does enabling `<think>` mode materially facilitate perception, or just the narrative? (Q6) Are there *fairness* gaps across subpopulations (e.g., gender, race, lighting, texture) under shifts?

For addressing the aforementioned research questions, we introduce The Perceptual Observatory , a holistic evaluation suite that measures *how* MLLMs see. We probe with (i) ID augmentations and (ii) OOD *stylized illusions* [18] images produced by diffusion with spatial control (Stable/Latent Diffusion + ControlNet) that alter appearance while preserving layout, letting us disentangle perception from priors [38, 54]. Tasks target complementary skills: identity matching (robustness to perturbations vs. distractors), grid pointing game (spatial invariance), and attribute localization for semi and fully guided settings [10] towards common-sense reasoning [11] assessment. We summarize our contributions as follows:

- We propose The Perceptual Observatory : A principled framework that evaluates perceptual robustness and vision-language grounding beyond tradition benchmark performance, highlighting whether failures stem from visual or textual capabilties.
- We consolidate simple, interpretable properties of MLLMs like identity robustness, spatial invariance, attribution fidelity, fairness gap, scale consistency, and effects of `<think>` mode to reveal *how* answers are grounded.
- To enable further research in this area, we also provide a scalable pipeline to generate ID corruptions and OOD *stylized illusions* (diffusion+ControlNet) that preserve spatial layout while confounding appearance.
- Finally, we provide a comprehensive analysis of three leading open-source MLLM families. We demonstrate that scaling the language model without proportional adaptation of the vision encoder results in systematic robustness gaps under distribution shifts, thereby pinpointing the methodological bottlenecks that future research must address.

2. Related Works

With the recent wave of MLLM families such as Qwen2.5-VL [43], Gemma3 [42], InternVL3.5 [51], etc., has dramatically pushed the boundaries of visual perception. The large-scale models have frozen or lightly adapted vision backbones such as ViT [8], SigLIP 2 [46], CLIP [36]. The early evaluation of these models has emphasized end-task accuracy. Benchmarks such as MMBench [32] and MMMU [52] extend text-centric evaluation to vision language understanding, offering huge collections of diverse QAs (c.f. MMLU [20]). Yet, these efforts lack perceptual understanding with language priors, leading to the question of whether the high scores on the benchmarks arise from the visual grounding or from the textual reasoning.

The computer vision community has long highlighted the fragility of models under distribution shifts [19, 21]. Analogous concerns have emerged for MLLMs. Experiments in abstract shape recognition show that VLMs often rely on texture or contextual clues rather than true shape understanding [18]. Similarly, [17, 39, 40, 55, 56] construct optical illusions and misleading visual scenarios. These works show that MLLMs are easily misled, as they capture end-task accuracy aided by prompting techniques to improve understanding, yet do not close the gap to human performance without explainability. Another flaw is that the models may have already been trained on certain popular illusions, such as Salvador Dali's painting [35]. QAs such as CLEVR [23] and Winoground [44] reveal that models fail to reason on spatial relations and subtle changes [47].

Beyond QA, VLMs may produce correct answers while attending to irrelevant regions, highlighting poor vision-language disentanglement. Thus, robust multimodal understanding requires attribution localization. Recent MLLMs predict bounding boxes for attributes, enabling explicit evaluation, but localization under distribution shifts for perturbations and illusion remains scarce.

While these prior benchmarks demonstrate critical weaknesses – language-prior exploitation, fragility to corruption, distribution shifts, and poor grounding – they traditionally examine **one dimension at a time**. The PERCEPTUAL OBSERVATORY fills this gap by providing a unified, property-driven assessment of MLLMs across robustness, grounding, and spatial reasoning with controlled low-level augmentations and high-level style-transfer illusions with tasks that explicitly measure identity preservation, spatial invariance, and attribution fidelity. Our OBSERVATORY yields a foundation for holistic insights for perceptual strength and weaknesses of MLLMs.

3. Perceptual Observatory

The PERCEPTUAL OBSERVATORY is a suite of assessments that characterizes multimodal LLMs across four axes: robustness, in-context adaptation, relational vision, and vision-language alignment as summarized in Figure 1. Unlike accuracy-only benchmarks, it examines *how* models perceive: whether they maintain identity under perturbations, transfer grounding across views, resist distractors, preserve spatial structure, or rely disproportionately on textual priors.

The framework is motivated by principles from perception and cognition, including feature integration [45] and structure mapping [14], which emphasize local-to-global organization and relational reasoning. We instantiate the Observatory in two canonical domains, face recognition and text-in-vision. Then expose models to controlled perturbations comprising (i) pixel-based augmentations (blur, jitter, noise, etc) and (ii) style-transfer based augmentations *"illusions"* generated via Diffusion [37]+ControlNet [54].

The PERCEPTUAL OBSERVATORY then evaluates parameter scales, and decoding modes across model families, yielding comprehensive perceptual profiles capturing robustness behavior, fairness gaps, vision-language alignment, and sensitivity to perturbations. These insights enable principled comparison and selection of reliable MLLM candidates.

3.1. Problem Statement

MLLM Characterization. We aim to evaluate how a pretrained multimodal LLM f behaves under controlled visual perturbations. Each sample in our dataset is a tuple (x, y, b), where x is an image, y is its label (identity or word), and b contains any available ground-truth attribute boxes. For a perturbation t drawn from a transformation set $\mathcal{T}$, the model is queried on the modified image $x' = t(x)$. For a given property P (e.g., identity matching, attribute localization), we collect the model's outputs relevant to that property and measure performance with a task-specific metric M. This formulation is task-agnostic and accommodates robustness, in-context adaptation, relational vision, and vision-language alignment.

Benchmark Datasets. We build two datasets with labeled attributes: (i) **CELEB** [1], a collection of celebrity faces with identity labels and bounding boxes for eyes, nose, and mouth; and (ii) **WORD**, a set of synthetically rendered "text" images with ground-truth labels and bounding boxes marking the text span.

Perturbations. Each dataset has two corresponding sets of perturbed images: (i) *Augmentations* ($\mathcal{T}_{\text{id}}$), consisting of 15 pixel-level transformations such as blur, jitter, and noise; and (ii) *Illusions* ($\mathcal{T}_{\text{ood}}$), 15 stylized transformations, which alter appearance while preserving spatial layout. For each image x, we sample a transformation t from either set to obtain $x' = t(x)$. The complete set of inputs considered in our evaluation is

$$\mathcal{T} = \{\text{Org}\} \cup \mathcal{T}_{\text{id}} \cup \mathcal{T}_{\text{ood}}$$

where Org denotes the unperturbed original image x.

3.2. In-Context Formulation

We frame all evaluations as in-context prediction. A model f is conditioned on a support example S an image together with a prompt (and, when relevant, text annotations) and must answer a query Q. Unless otherwise specified, the support is the original image $x^{(\text{Org})}$.

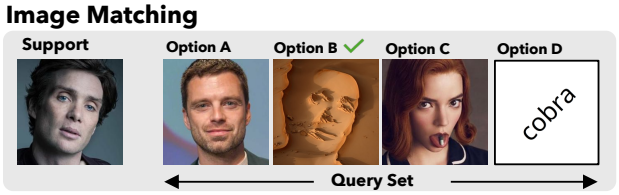

Figure 2. Image Matching: the model selects the candidate, matching the support image. (Supp Sec: Prompt A Prompt B, Prompt C

[1] HF Dataset

Task 1: Image Matching. The model is shown a support image and must choose which element in a four-way query set depicts the same entity. As illustrated in Figure 2 (see *Image Matching*), the query set contains four images arranged as *Option A–D* in the figure:

1) the **correct match** (*option B* in figure), a perturbed version of the support entity (e.g., blurred, stylized, or otherwise transformed); 2) an **out-of-context** sample drawn from the other domain (face vs. text) (*option D*); 3) two **distractors** (*option A & C*), chosen as near neighbors CLIP-based nearest faces or words with ±1 character edits).

Given support S and candidates $\{\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}\}$, the model must output the correct option choice.

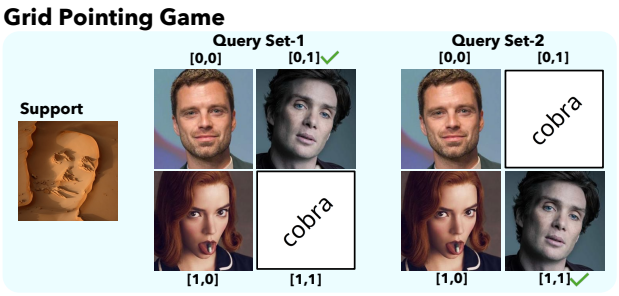

Figure 3. Grid Pointing Game: the model identifies the grid position containing the Org image. (Supp Sec: Prompt D Prompt E, Prompt F

Task 2: Grid Pointing Game. The model is given a support image and a 2×2 collage (not limitated to) in which the original image $x_e^{(\mathrm{Org})}$ (*correct option* `[0,1]` in query set-1 and `[1,1]` query set-2) is placed at one of four positions ℓ; the remaining three cells contain distractors or out-of-context samples (constructed as in Task 1). As shown in Figure 3, the model must point to the location containing the original image by predicting $\hat{\ell}$. Each entity appears once in every grid position across query sets.

Task 3: Attribute Localization. For an entity e with attributes $\mathcal{A}_e$ and ground-truth boxes $\{b_{e,a}\}$, the model is given a support image (with one or more annotated boxes) and must predict the corresponding attribute boxes $\hat{b}_{e,a}(t)$ on a perturbed query image $x_e^{(t)}$. As shown in Figure 4, the task evaluates how well the model preserves spatial and structural information under appearance changes.

We consider two variants:

a **Semi-guided (one-hint):** the support provides a single attribute box, and the model must infer the remaining attributes, probing spatial commonsense.

b **Guided (full-hints):** the support provides all attribute boxes, and the model must transfer them to perturbed views probing perceptual consistency.

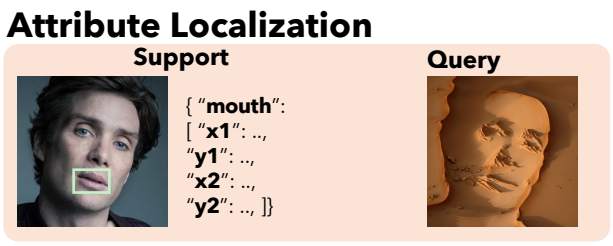

Figure 4. Attribute Localization: the model has to identify attribute information from the support image to the perturbed query. Semi-guided: (Supp Sec: Prompt J, Prompt L); Guided: (Supp Sec: Prompt I Prompt H, Prompt K, Prompt G)

3.3. Properties

We evaluate both the perceptual robustness of MLLMs and their vision-language alignment. Each property corresponds to an intuitive behavioral goal and a simple quantitative metric.

Identity Matching Robustness. Used for Image Matching and Grid Pointing Game across both datasets. A robust model should preserve entity identity under id and OOD perturbations. We measure the accuracy drop $\Delta = \mathrm{Acc}(\mathrm{x}^{\mathrm{Org}}) - \mathrm{Acc}(x^t)$, where $t \sim \mathcal{T}_{\mathrm{id}} \cup \mathcal{T}_{\mathrm{ood}}$. Smaller values indicate stronger identity tracking.

Gender Bias. Evaluated on CELEB for all tasks. A fair model should perform similarly on male and female identities. We compute GAP $= Z_M - Z_F$, using IoU or accuracy depending on the task. Low magnitude of GAP indicates gender-neutral behavior.

Invariance to Spatial Arrangements. Specific to the Grid Pointing Game. A position-invariant model should not rely on the grid location of the correct image. For per-position accuracies Acc^{ℓ}, we report $\mathrm{Gap}_{\ell} = \max_{\ell} \mathrm{Acc}^{\ell} - \min_{\ell} \mathrm{Acc}^{\ell}$. Smaller spreads reflect stronger spatial invariance.

Scale Consistency. Evaluated across all tasks and datasets. As model size increases within a family, scores Z_k should improve monotonically with parameter count N_k. We summarize the average gain per parameter doubling. Positive trends indicate scalable perceptual grounding.

Thinking Superiority. Evaluated across all tasks and datasets. Reasoning-enabled decoding (`<think>` mode) should enhance perceptual performance. For matched settings, we compute $\Delta^{\mathrm{think}} = Z^{\texttt{<think>}} -$

$Z^{\texttt{base}}$. Positive values indicate that chain-of-thought decoding benefits recognition and grounding.

Salient Perceptual Understanding. Used for Attribute Localization (Task 3). A strong model should preserve salient structure when localizing attributes. *(a) Semi-guided:* we measure the gain from providing one hint, probing spatial commonsense. *(b) Guided:* we evaluate transfer retention (TR),

$$\mathrm{TR}(t) = \frac{\mathrm{mIoU}_{\mathrm{guided}}(t)}{\mathrm{mIoU}_{\mathrm{guided}}(\mathrm{Org})},$$

which tests whether full supervision transfers to perturbed views. High TR indicates stable perceptual layouts under id and OOD shifts.

4. Experiments

4.1. Dataset

We construct a two-part benchmark to probe perceptual abilities of multimodal LLMs (MLLMs). **CELEB** contains 1,000 celebrity face images with gold bounding boxes for key features (eyes, nose, mouth), derived from `MediaPipe` [34] and authors manually verified 10% of the samples and achieved 98% IoU w.r.t gold. **WORD** consists of ∼267K procedurally rendered words across 21 semantic categories, rendered under diverse fonts, casings, positions, and rotations, yielding >1M unique images with exact ground-truth bounding boxes.

To study robustness, we apply two perturbation families: (1) $\mathcal{T}_{\mathrm{id}}$ - content-preserving linear augmentations (using Albumentations [6]), and (2) $\mathcal{T}_{\mathrm{ood}}$ - style/illusion perturbations using ControlNet [54] and Stable Diffusion [37]. Each image has 15 $\mathcal{T}_{\mathrm{id}}$ variants, 15 $\mathcal{T}_{\mathrm{ood}}$ variants, and the original, yielding 31K images per dataset and 62K in total.

Further implementation details (augmentation lists, prompt templates, scaling factors) are provided in the supplementary material.

4.2. Implementation

MLLMs setup. We use a variety of MLLMs, including 3 distinct model families: (1) `Qwen2.5-VL-(3B/7B/72B)-Instruct` [4, 43, 50], (2) `Gemma-3-(4B/12B/27B)-Instruct` [42], and (3) `InternVL3.5`[2]`-(8B/14B)-(Instruct/Thinking)` [51]. The selection was strategically designed to cover a broad spectrum and avoid single evaluation. The key factors included a suite of parameter sizes, distinct model architectures, reasoning capabilities, multi-image inputs, and date of release. All experiments were conducted on HPC clusters equipped with NVIDIA 4×H200s with 144GB and 4×H100s with 80GB VRAM, utilizing PyTorch, Huggingface, and the vLLM [24] framework. We maintained a constant `temperature` of 0.2, `top_p` of 0.95, and `top_k` of 32 throughout our experimentation.

[2]HF Transformer compatible

5. Results & Discussion

As a preview of our results that we will describe in detail, we establish four consistent themes across properties defined in §3.3: (1) **WORD** tasks are near-saturated in-distribution and retain accuracy under shift, whereas **CELEB** tasks degrade sharply out-of-distribution. (2) Scaling primarily benefits OCR, pointing, and guided localization, but does not guarantee robustness to identity-preserving perturbations. (3) **LM-side capacity** (decoder depth/width, projector dimension) drives most of the gains, since the vision encoder is held fixed. (4) Decode-time reasoning (`<think>`) enhances clean/ID performance but reduces transfer retention on faces. Humans achieve near-ceiling accuracies, highlighting that gaps are model-driven rather than dataset artifacts.

5.1. Identity Matching Robustness

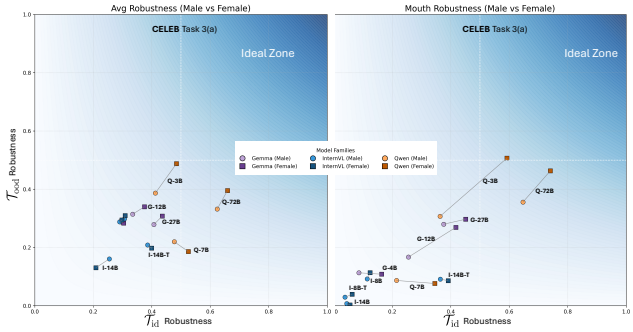

Figure 5. Left figure demonstrates Multidimensional Insights for Task 3(a) across all attributes (eyes, nose, mouth), gender gap, robustness on ID vs. OOD for CELEB . Whereas right figure provides fine-grain insights for a specific attribute: `mouth`.

Table 1 depicts Δ under $\mathcal{T}_{\mathrm{id}}$ and $\mathcal{T}_{\mathrm{ood}}$. Three robust trends emerge: (i) ID augmentations (blur, noise, etc.) produce negligible loss and occasionally improve accuracy. (ii) OOD illusions disproportionately harm *mid-scale* models (7–14B), along with larger MLLMs (Qwen-72B, Gemma-27B) fail to retain higher robustness. (iii) Robustness is non-monotonic: e.g., Gemma-12B is more brittle than Gemma-4B, highlighting that methodological flaws can outweigh scale. Grounding in Table 1 similar trends can be inferred for WORD across Org/ID/OOD. Human annotators exceed 95%

Table 1. Table summarizes robustness of MLLMs for ID vs OOD across both dataset across all task, here Task3(b) is dubbed as Task 3. Δ refers to difference between current vs smallest among family ex: Qwen7B – Qwen3B, and also difference between thinking (<T>) vs non-thinking.

	CELEB									WORD								
	Task1			Task2			Task3			Task1			Task2			Task3		
#	Org	$\mathcal{T}_{id}$	$\mathcal{T}_{ood}$	Org	$\mathcal{T}_{id}$	$\mathcal{T}_{ood}$	Org	$\mathcal{T}_{id}$	$\mathcal{T}_{ood}$	Org	$\mathcal{T}_{id}$	$\mathcal{T}_{ood}$	Org	$\mathcal{T}_{id}$	$\mathcal{T}_{ood}$	Org	$\mathcal{T}_{id}$	$\mathcal{T}_{ood}$
Qwen 2.5 - VL																		
3B	33.66	29.57	22.57	25.00	24.95	24.98	90.57	90.23	**85.33**	21.00	25.66	24.66	25.75	25.53	25.08	97.54	97.48	**95.35**
7B	**78.21**	**80.52**	10.00	64.75	65.66	29.81	99.65	99.17	16.52	**75.00**	**76.26**	19.33	43.75	47.43	36.33	**99.99**	**99.95**	54.39
Δ	44.55	50.95	-12.57	39.75	40.71	04.83	09.08	08.94	-68.81	54.00	50.60	-05.33	18.00	21.90	11.25	02.45	02.47	-40.96
72B	51.48	65.10	**31.61**	**98.25**	**98.66**	**48.40**	**99.99**	**99.93**	39.65	56.00	64.06	**36.80**	**87.75**	**88.40**	**47.06**	98.64	97.97	50.55
Δ	17.82	35.53	09.04	73.25	73.71	23.42	09.42	09.70	-45.68	35.00	38.40	12.14	62.00	62.87	21.98	01.10	00.49	-44.80
Gemma 3																		
4B	24.75	30.09	25.67	45.00	47.10	28.85	**99.83**	**99.72**	**61.92**	23.00	27.66	29.20	46.25	45.56	29.56	99.99	99.99	**99.69**
12B	55.44	67.98	32.34	**84.75**	**84.44**	**40.03**	99.54	98.47	51.86	66.00	65.46	30.53	64.50	64.93	**40.68**	99.99	99.99	99.28
Δ	30.69	37.89	06.67	39.75	37.34	11.18	-00.29	-01.25	-10.06	43.00	37.80	01.33	18.25	19.37	11.12	00.00	00.00	-00.41
27B	**78.21**	**84.35**	**43.16**	71.50	68.63	31.40	99.58	97.38	25.86	**96.00**	**93.20**	**32.13**	**69.00**	**67.50**	38.46	**99.99**	**99.99**	80.27
Δ	53.46	54.26	17.49	26.50	21.53	02.55	-00.25	-02.34	-36.06	73.00	65.54	02.93	22.75	21.94	08.90	00.00	00.00	-19.42
InternVL 3.5																		
8B	18.81	23.56	26.86	81.75	79.71	36.56	**99.99**	**99.99**	**66.84**	13.00	19.11	26.20	47.00	44.53	27.31	**99.99**	**99.99**	**99.64**
– <T>	03.96	13.53	15.97	29.25	31.73	26.66	96.99	95.50	17.79	08.00	12.99	13.40	06.25	06.33	09.56	72.99	66.76	47.82
Δ	-14.85	-10.03	-10.89	-52.50	-47.98	-09.90	-03.00	-04.49	-49.05	-05.00	-06.12	-12.80	-40.75	-38.20	-17.75	-27.00	-33.23	-51.82
14B	27.72	**40.66**	26.66	**98.50**	**97.91**	47.08	49.41	53.39	52.34	**34.00**	**37.86**	26.73	**71.25**	**72.40**	43.55	98.61	98.12	95.42
– <T>	**27.72**	35.77	**27.19**	89.25	92.53	**51.18**	85.04	81.49	11.17	27.00	33.13	**28.46**	44.00	44.59	**49.41**	99.22	97.30	92.89
Δ	00.00	-04.89	00.53	-09.25	-05.38	04.10	35.63	28.10	-41.17	-07.00	-04.73	01.73	-27.25	-27.81	05.86	00.61	-00.82	-02.53
Human																		
	100	100	89.11	100	100	87.55	95.66	93.22	81.88	100	100	83.86	100	100	79.98	99.99	99.99	99.99

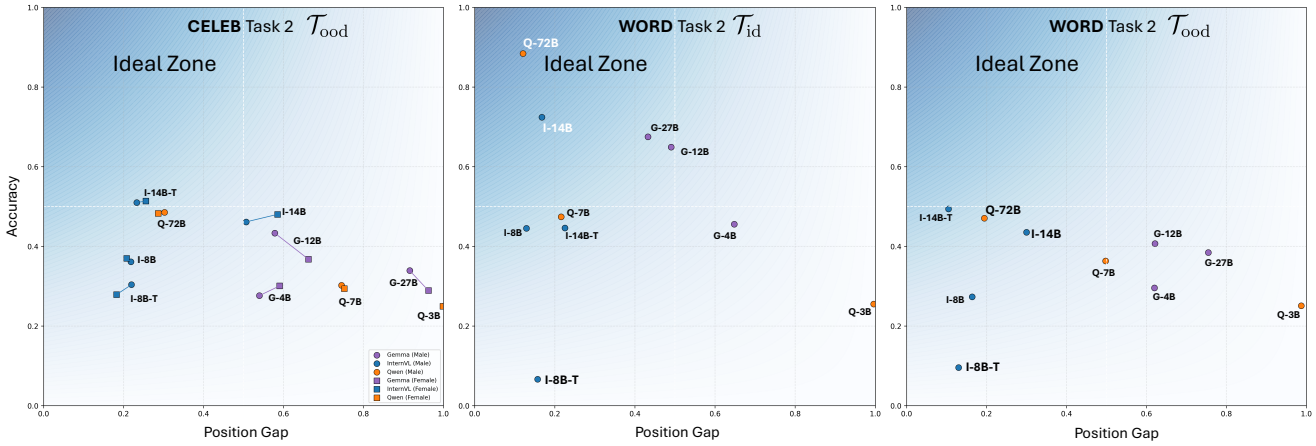

Figure 6. Multidimensional insights for Task 2, accuracy vs. position gap across perturbation, across datasets, this figure reveals majority of the models suffer under OOD setting with high gender gap suggesting sensitivity to grid position.

across all conditions, establishing an empirical ceiling.

5.2. Invariance to Spatial Arrangements

In the Grid Pointing Game, Figure 6 reveals insights that several models show pronounced positional biases, with gap spreads (§ 3.3) exceeding *approx.* 50 – 90% for small–to–large MLLMs. For CELEB, only `InternVL-3.5-8B-thinking` has the least position bias, but also has the worst accuracy, depicting that, *thinking* doesn't facilitate in *seeing*. Even for simple WORD, models (`Qwen2.5-VL-72B` & `InternVL-3.5-14B`) which seem to be in "ideal zone" tend to fail as we switch from ID to OOD this suggest stylistic change were not incorporated by language understanding, as vision encoder was never aligned jointly. Larger decoders reduce Gap_ℓ on WORD but only partially on CELEB, confirming that the encoder, regulate spatial invariance.

5.3. Gender Bias in CELEB

Figure 5 (a) and Figure 6 provide a visualization to understand gender bias (depicted as line between round

and square marker; *larger line depicts huge gender bias*) in MLLMs along with assessing other axes like robustness for OOD vs. ID. As depicted in Figure 6, models like `InternVL` and `Gemma` have gender bias, especially for `Gemma-12B` where-in the model goes from low position gap to worst when gender is changed from male to female.

One can observe in Figure 5 that for Task 3(a): Semi-guided attribution task, most models have gender bias, when analyzing with fine-grain lens, example: just for Attribute: `mouth`; models become much worse (Ex: `Qwen2.5-VL-3B` & `72B`, `Gemma-12B`). Due to the visual backbone being unchanged across sizes, we hypothesize that, gains arise from stronger cross-modal calibration in the textual space rather than visual. Nonetheless, asymmetries persist without explicit debiasing.

5.4. Scale Consistency

Table 1 summarizes scale-consistency, i.e. just scaling the language model may not be the right way to improve performance, as `Qwen`(`3B`→`7B`), `InternVL`(`8B`→`14B`) , and `Gemma`(`4B`→`12B`→`27B`) performance is improved drastically on Task 1: (Org, $\mathcal{T}_{\text{id}}$) for CELEB & WORD, whereas further scaling collapses the `Qwen` to `72B` performance. Contrary, for Task 1: ($\mathcal{T}_{\text{ood}}$) the model performance is either half of the Org, $\mathcal{T}_{\text{id}}$; or its smaller variant. This suggests that the models rely on the training *"world knowledge"* rather than focusing on *"visual cues"*. This clearly necessitates the need for joint alignment of both vision-encoder and language-decoder for scaling.

5.5. Task-Level Grounding

Task-3 (Attribute Localization). Figure 5 Task 3(a) shows no models are even close to the Ideal Zone (High IoU for $\mathcal{T}_{\text{id}}$, $\mathcal{T}_{\text{ood}}$). Even from Table 1 Task 3(b), a simply cognitive task of attribute transcription, larger models perform poorly compared to smaller counter parts. **Mouth Localization**. Figure 5 highlights that ID distributions peak at high IoU, but OOD shifts the distribution with a low-IoU for almost all models, but especially all `InternVL` variants suffer from OOD distribution change. This corroborates our spatial-invariance findings.

5.6. Thinking Superiority

We assess, whether `<think>` mode actually thinks? Table 1 Task 1,2,3(b), `InternVL-3.5-8B-thinking` fails on all tasks compared to its non-thinking variant, as highlighted in red color. `InternVL-3.5-14B-thinking` also follows similar trends of poor performance compared to the non-thinking variant. Figure 5 shows that for attribute: mouth, the thinking variant of

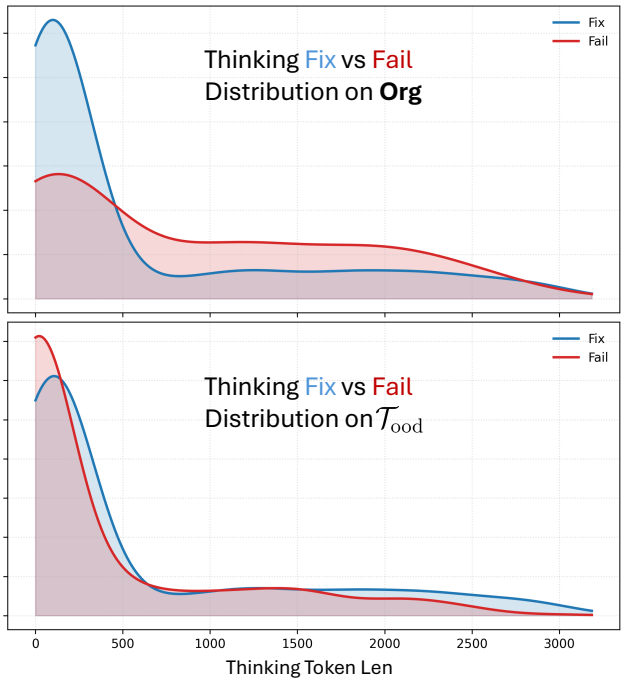

Figure 7. Celeb chain length vs. outcome. Histogram (log-y) of `<think>` token length for cases where reasoning *fixes* vs. *fails*. Top: Org fixes vs. fails. Bottom: $\mathcal{T}_{\text{ood}}$ fixes vs. fails.

`InternVL` has the lowest robustness compared to other non-thinking models. Moreover, Figure 6 `InternVL-3.5-8B-thinking` has the lowest position-gap for WORD, but of no use as the accuracy is very poor (below 10%). Also, for CELEB, it has more gender-bias compared to its non-thinking counterpart.

Reasoning chain length. Figure 7 shows that on Org, successful fixes concentrate at short reasoning chain lengths, while failures still do occur when operating with chain length $\sim$ 2000 tokens. Contrary, for $\mathcal{T}_{\text{ood}}$, rare fixes appear in longer-chain tails, and the model most of the time gives up during the early stage of reasoning with high confidence, suggesting over-reliance on textual knowledge compared to visual.

5.7. Salient Perceptual Understanding

From Figure 5, we observe that on Task 3(a), almost all models suffer in spatial-common sense understanding, i.e, given "nose" or "top-left-corner" coords, models struggle to identify other attributes which are spatially very near. Transfer retention drops for simple cognitive tasks like guided transcription and semi-guided attribute localization on CELEB & WORD for $\mathcal{T}_{\text{ood}}$. We demonstrate other multi-dimensional vulner-

abilities like gender-bias, spatial-invariance, robustness to OOD samples, scaling effects, and true performance of `<think>` mode of current MLLMs using Figure 6 & Figure 5.

5.8. Human Baseline

To contextualize model performance, we conducted a human study on both **CELEB** and **WORD**. For each dataset, 100 samples were randomly chosen and evaluated across all three tasks (§3.2) by two annotators, achieving an average inter-annotator agreement of 94.5%.

As shown in Table 1, humans achieved near-perfect accuracies ($> 95\%$) on identity and spatial tasks, with only mild degradation under $\mathcal{T}_{\text{ood}}$ perturbations. On attribute localization, annotators retained high performance (81% mIoU in the most challenging guided-perturbation setting), even in the semi-guided case.

These results establish the empirical upper bound: the tasks are perceptually tractable for humans, and gaps in robustness, spatial invariance, or grounding can be attributed to limitations of current MLLMs rather than dataset artifacts.

6. Limitations

While the Perceptual Observatory provides a principled framework for assessing MLLMs, several limitations remain. First, the evaluation is restricted to two domains (faces and synthetic words), limiting conclusions about broader perceptual generalization. Second, human annotations for illusions were verified only on a subset, and baselines were derived from a small sample with few annotators, which constrains statistical robustness. Third, fairness analysis focused on gender, leaving other social factors such as skin tone unexplored. Fourth, experiments were limited to open-source models for transparency and feasibility, excluding closed-source systems. These choices were deliberate to ensure tractability and interpretability, but expanding datasets, annotations, social dimensions, and model coverage remains an important direction for future work.

7. Conclusion & Future Work

This work introduced The Perceptual Observatory , a principled framework for holistic evaluation of visual capabilities of MLLMs, by combining controlled pixel-based augmentations along with diffusion-based styled illusions, and by evaluating tasks spanning identity matching, grid-based spatial reasoning, and attribute localization. This Observatory moves beyond traditional leaderboard benchmarks. Our proposed property and insights lay the foundation for robustness, failures arising from vision encoders, language decoder scaling, and reasoning capabilities that reveals inherent flaws in grounding and fairness across model families and sizes. We observed, scaling language decoders does not guarantee monotonic gains in visual grounding and hinders the visual understanding under OOD distribution shifts. These insights showcase the importance of evaluating how models "see", not only how well they answer, and provide actionable insights for designing next-generation multimodal models.

To extend the impact of the Perceptual Observatory , we aim to broaden the dataset scope beyond celebrity faces and synthetic words to more diverse visual domains. This will enable more comprehensive and holistic evaluation of multimodal models' visual strengths and weaknesses. Furthermore, the expanded dataset will serve as a foundation for joint vision-language alignment. Instead of scaling only the language component, we propose a joint optimization framework that simultaneously scales both vision and language components. Leveraging reinforcement learning for post-training, we will use the property-based metrics defined in this work as rewards. This approach ensures that vision is given equal importance, potentially improving alignment and robustness.

Additionally, we identify the need for a deeper evaluation of reasoning chains in MLLMs. While our analysis touched on reasoning-enabled decoding, there is currently no standard metric for evaluating reasoning quality. Future work will develop and incorporate such metrics to provide a clearer understanding of how reasoning chains contribute to model performance and robustness.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[2] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 1, 2

[4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang,

Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv preprint arXiv:2308.12966*, 2023. 5

[5] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. 2

[6] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. Albumentations: Fast and Flexible Image Augmentations. *Information*, 11(2), 2020. 5, 12

[7] Wenliang Dai, Junnan Li, DONGXU LI, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In *Advances in Neural Information Processing Systems*, pages 49250–49267. Curran Associates, Inc., 2023. 1

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[9] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. PaLM-E: an embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*. JMLR.org, 2023. 1

[10] Bryan Etzine, Masoud Hashemi, Nishanth Madhusudhan, Sagar Davasam, Roshnee Sharma, Sathwik Tejaswi Madhusudhan, and Vikas Yadav. Revitalizing saturated benchmarks: A weighted metric approach for differentiating large language model performance. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 511–523, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. 2

[11] Xingyu Fu, Muyu He, Yujie Lu, William Yang Wang, and Dan Roth. Commonsense-T2I challenge: Can text-to-image generation models understand commonsense? *arXiv preprint arXiv:2406.07546*, 2024. 2

[12] Robert Geirhos, Carlos R. M. Temme, Jonas Rauber, Heiko H. Schütt, Matthias Bethge, and Felix A. Wichmann. Generalisation in humans and deep neural networks. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2018. 2

[13] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. 2

[14] Dedre Gentner. Structure-Mapping: A Theoretical Framework for Analogy. *Cognitive Science*, 7(2):155–170, 1983. 3

[15] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 2

[16] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. 1

[17] Tianyang Han, Qing Lian, Rui Pan, Renjie Pi, Jipeng Zhang, Shizhe Diao, Yong Lin, and Tong Zhang. The instinctive bias: Spurious images lead to illusion in MLLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16163–16177, Miami, Florida, USA, 2024. Association for Computational Linguistics. 2

[18] Arshia Hemmat, Adam Davies, Tom A. Lamb, Jianhao Yuan, Philip Torr, Ashkan Khakzar, and Francesco Pinto. Hidden in Plain Sight: Evaluating Abstract Shape Recognition in Vision-Language Models. In *Advances in Neural Information Processing Systems*, pages 88527–88556. Curran Associates, Inc., 2024. 2, 12

[19] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 2

[20] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020. 2

[21] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8340–8349, 2021. 2

[22] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15262–15271, 2021. 2

[23] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross

Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 2

[24] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023. 5

[25] Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253, 2017. 2

[26] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. SEED-Bench-2: Benchmarking Multimodal Large Language Models. *arXiv preprint arXiv:2311.17092*, 2023. 1

[27] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

[28] Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. SEED-Bench-2-Plus: Benchmarking Multimodal Large Language Models with Text-Rich Visual Comprehension. *arXiv preprint arXiv:2404.16790*, 2024. 1

[29] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1

[30] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong 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, Singapore, 2023. Association for Computational Linguistics. 1

[31] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1, 2

[32] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024. 1, 2

[33] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts. In *The Twelfth International Conference on Learning Representations*, 2024. 1

[34] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019. 5

[35] Susana Martinez-Conde, Dave Conley, Hank Hine, Joan Kropf, Peter Tush, Andrea Ayala, and Stephen L. Macknik. Marvels of Illusion: Illusion and Perception in the Art of Salvador Dalí. *Frontiers in Human Neuroscience*, 9:496, 2015. 2

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 3, 5, 12

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn 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. 2

[39] Mohammadmostafa Rostamkhani, Baktash Ansari, Hoorieh Sabzevari, Farzan Rahmani, and Sauleh Eetemadi. Illusory VQA: Benchmarking and enhancing multimodal models on visual Illusions. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2995–3004, 2025. 2

[40] Haz Sameen Shahgir, Khondker Salman Sayeed, Abhik Bhattacharjee, Wasi Uddin Ahmad, Yue Dong, and Rifat Shahriyar. Illusionvqa: A challenging optical illusion dataset for vision language models. *arXiv preprint arXiv:2403.15952*, 2024. 2

[41] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA Models That Can Read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[42] Gemma Team. Gemma 3. 2025. 2, 5

[43] Qwen Team. Qwen2.5-VL, 2025. 2, 5

[44] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022. 2

[45] Anne Treisman and Garry A. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12:97–136, 1980. 3

[46] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 2

[47] Aayush Atul Verma, Amir Saeidi, Shamanthak Hegde, Ajay Therala, Fenil Denish Bardoliya, Nagaraju Machavarapu, Shri Ajay Kumar Ravindhiran, Srija Malyala, Agneet Chatterjee, Yezhou Yang, et al. Evaluating multimodal large language models across distribution shifts and augmentations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5314–5324, 2024. 2

[48] Johan Wagemans, James H Elder, Michael Kubovy, Stephen E Palmer, Mary A Peterson, Manish Singh, and Rüdiger Von der Heydt. A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure–ground organization. *Psychological bulletin*, 138(6):1172, 2012. 2

[49] Johan Wagemans, Jacob Feldman, Sergei Gepshtein, Ruth Kimchi, James R Pomerantz, Peter A Van der Helm, and Cees Van Leeuwen. A century of Gestalt psychology in visual perception: II. Conceptual and theoretical foundations. *Psychological bulletin*, 138(6): 1218, 2012. 2

[50] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191*, 2024. 5

[51] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 2, 5

[52] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 2

[53] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024. 1

[54] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3836–3847, 2023. 2, 3, 5, 12

[55] Yichi Zhang, Jiayi Pan, Yuchen Zhou, Rui Pan, and Joyce Chai. Grounding visual illusions in language: Do vision-language models perceive illusions like humans? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5718–5728, Singapore, 2023. Association for Computational Linguistics. 2

[56] Yiming Zhang, Zicheng Zhang, Xinyi Wei, Xiaohong Liu, Guangtao Zhai, and Xiongkuo Min. IllusionBench: A Large-scale and Comprehensive Benchmark for Visual Illusion Understanding in Vision-Language Models. *arXiv preprint arXiv:2501.00848*, 2025. 2

The Perceptual Observatory
Characterizing Robustness and Grounding in MLLMs

Supplementary Material

8. Dataset Details

8.1. CELEB

We sample 1,000 celebrity face images for facial feature attribution. Bounding boxes for left/right eyes, nose, and mouth are computed using `MediaPipe`. To validate reliability, the first and second authors manually annotated 10% of images, achieving 98% IoU with `MediaPipe` outputs. Hence, we treat `MediaPipe`-derived boxes as gold annotations.

8.2. WORD

We collect ∼267K unique words across 21 semantic categories (*Computer Science, Cities, People, Food, Politics, Abuse*, etc.). Word length $l \in [2, 10]$ with $\mathbb{E}[l] \approx 4.8$. Each word is rendered under:

$$\mathcal{F} \times \mathcal{C} \times \mathcal{P} \times \mathcal{R},$$

with $\mathcal{F} = \{CourierNew, ..., TimesNewRoman\}$ (fonts), $\mathcal{C} = \{\text{upper}, \text{lower}, \text{camel}\}$ (casings), $\mathcal{P} = \{\text{center}, \text{top}, \text{bottom}\}$ (positions), $\mathcal{R} = \{-45°, 0°, 45°\}$ (rotations). Uniform sampling across these factors produces >1M rendered images overall. Because WORD is procedurally generated, bounding boxes are exactly known.

8.3. Perturbations

We apply two perturbation families:

Linear augmentations ($\mathcal{P}_1$). Implemented with Albumentations [6]. Each image is augmented by sampling from the set

$$\mathcal{M} = \left\{ \begin{array}{l} \texttt{GaussianBlur}(11, 11),\ \texttt{MedianFilter}(21), \\ \texttt{ZoomBlur}([1.05, 1.07]),\ \texttt{ChromaticAberration}(\pm 0.2), \\ \texttt{ISONoise}([0.01, 0.05], [0.1, 0.5]),\ \texttt{RGBShift}(\pm 20), \\ \texttt{Salt\&PepperNoise}([10^{-4}, 10^{-3}]),\ \texttt{GammaLimit}([80, 140]), \\ \texttt{JPEGCompression}([20, 50]),\ \texttt{MultiplicativeNoise}([0.9, 1.1]), \\ \texttt{Sharpen}(\alpha \in [0.3, 0.5]),\ \texttt{GlassBlur}(\sigma = 0.3, \Delta = 2), \\ \texttt{Posterize}(4 \text{ bits}),\ \texttt{MotionBlur}(7, 7), \\ \texttt{GaussianNoise}(\mu = 0,\ \sigma \in [0.05, 0.1]) \end{array} \right\}$$

Thus, $\mathcal{P}_1(x) \sim \mathcal{U}(\mathcal{M})$.

Illusion perturbations ($\mathcal{P}_2$). Following IllusionBench [18], each source image x_i is embedded into a stylized scene using ControlNet [54] with Stable Diffusion [37]. Prompts are composed from:

$[\texttt{SubjectScene}] \times [\texttt{Style}] \times [\texttt{Light/ColorHighlight}]$,

where representative values are listed below:

Subject Scene	Style	Light/Color Highlight
Museum	Cinematic	Dust Motes
Rainy Alleyway	Gothic	Neon Glow
Forest	Fantasy Art	Golden Hour
Desert Dune	Vintage Photo	Pastel Hues
Medieval Village	Minimalist	Stark Shadows
Ocean	Surrealism	Electric Blue
Sunset Beach	Bioluminescent	Crystal Refraction
Cozy Cottage	Origami	Venetian Blinds
Mountain Range	Dystopian	Hearth Fire
Overgrown Ruins	Abstract	Volumetric Rays
Starry Night	Painting	Smudged Grays
Cloudy	Pixel Art	Pink Cyan

We apply a negative prompt (`glitch, low quality`) to suppress artifacts. Control strengths are dataset-dependent: WORD: $cn\_scale = 1.2$, $guide\_scale = 10.5$, CELEB: $cn\_scale = 3.0$, $guide\_scale = 7.5$.

Each final entry is stored as (x_{ij}, s_j), where s_j encodes the sampled scene, style, and lighting.

Final Dataset Size

For each dataset (CELEB, WORD), we sample 1,000 original images and generate 15 variants with $\mathcal{P}_1$, 15 with $\mathcal{P}_2$, plus the original. This yields:

$$1000 \times (1 + 15 + 15) = 31{,}000 \text{ images per dataset.}$$

In total, the benchmark contains **62,000** images.

9. Prompt Templates

Prompt A: Image Matching Query

```
**INSTRUCTIONS**
You are given 4 images each of size 1024x1024.

**TASK**
Compare the support image with 4 candidate images, and
    ↪ select a single candidate image that best matches
    ↪ the support image.

Return the result as valid JSON with detailed reasoning.

**JSON output format**
‘‘‘json
{
  "reasoning": "Provide a structured explanation based on
      ↪ visual cues. Cite concrete visual evidence and
      ↪ justify your identification.",
  "final_answer": "A" or "B" or "C" or "D"
}
‘‘‘
```

Prompt B: Image Matching Support (Celeb)

```
**CONTEXT**
You are given an image of size 1024x1024 of a famous person
    ↪ . This information serves as a factual reference.
Additionally, you will further be given 4 images, only one
    ↪ of them is generated from this image, where a face
    ↪ might be clearly visible, perturbed, stylized, or
    ↪ blended with the environment (background or object)
    ↪  as a visual illusion.
```

Prompt C: Image Matching Support (Word)

```
**CONTEXT**
You are given an image of size 1024x1024 of a case
    ↪ sensitive sequence of characters, "[WORD_LABEL]".
    ↪ This information serves as a factual reference.
Additionally, you will further be given 4 images, only one
    ↪ of them is generated from this image, where a
    ↪ sequence of characters are clearly written,
    ↪ perturbed, stylized, or blended with the
    ↪ environment (background or object) as a visual
    ↪ illusion.
```

Prompt D: GPG Query

```
**INSTRUCTIONS**
You are given an image of size 1024x1024 that is composed
    ↪ of 4 sub-images arranged in a 2x2 grid as a collage
    ↪ .
Only one of these sub-images is the *source image* from
    ↪ which the support image was generated.

**TASK**
Identify and locate which grid cell contains the source
    ↪ image.
Coordinates mapping:
[0,0] = top-left
[0,1] = top-right
[1,0] = bottom-left
[1,1] = bottom-right

Return the result as valid JSON with detailed reasoning.

**JSON output format**
‘‘‘json
{
  "reasoning": "Provide a structured explanation based on
    ↪ visual cues. Cite concrete visual evidence and
    ↪ justify your identification.",
  "final_answer": "[0,0]" or "[0,1]" or "[1,0]" or "[1,1]"
}
‘‘‘
```

Prompt E: GPG Support (Celeb)

```
**CONTEXT**
You are given an image of size 1024x1024. This image is a
    ↪ visually altered version of some original image
    ↪ such that the source might contain a face of famous
    ↪  person that have been clearly visible, perturbed,
    ↪ stylized, or blended with the environment (
    ↪ background or object) as a visual illusion.
The information serves as the support context.
```

Prompt F: GPG Support (Word)

```
**CONTEXT**
You are given an image of size 1024x1024. This image is a
    ↪ visually altered version of some original image
    ↪ such that the source might contain a sequence of
    ↪ characters that are clearly written, perturbed,
    ↪ stylized, or blended with the environment (
    ↪ background or object) as a visual illusion.
The information serves as the support context.
```

Prompt G: Attribution Support (Celeb)

```
**CONTEXT**
You are given an image of size 1024x1024 of a famous person
    ↪ , "[CELEB_LABEL]". The following text provides
    ↪ context for the key features present in this image,
    ↪  listing each attribute with its precise bounding
    ↪ box coordinates. This information serves as a
    ↪ factual reference.

Attributes:
‘‘‘json
[BBOX]
‘‘‘
```

Prompt H: Attribution Support (Word)

```
**CONTEXT**
You are given an image of size 1024x1024 of a case
    ↪ sensitive sequence of characters, "[WORD_LABEL]".
    ↪ The following text provides context for the
    ↪ sequence visible in this image, defining the
    ↪ characters and their precise bounding box. This
    ↪ information serves as a factual reference.

Attributes:
‘‘‘json
[BBOX]
‘‘‘
```

Prompt I: Attribution Guided Query (Word)

```
**INSTRUCTIONS**
You are given an image of size 1024x1024 that was generated
      ↪  from the above support image. The image contains a
      ↪  sequence of characters clearly written, distorted,
      ↪  stylized, or blended with the environment (
      ↪ background or object) as a visual illusion.

**TASK**
Think and analyze the image carefully. Using the support
      ↪ context as a reference, detect the sequence of
      ↪ characters and provide a single bounding box that
      ↪ encloses all of its characters with detailed
      ↪ reasoning.

Return the result as a valid JSON list. If no sequence is
      ↪ confidently located, return empty list [].

**JSON output format**
‘‘‘json
{
    "sequence": "The sequence of characters you read and
     ↪ detected",
    "reasoning": "Explain the visual evidence for this
     ↪ detection.",
    "x1": int,
    "y1": int,
    "x2": int,
    "y2": int
}
‘‘‘
```

Prompt J: Attribution Semi Guided Query (Word)

```
**INSTRUCTIONS**
You are given an image of size 1024x1024 that was generated
      ↪  from the above support image. The image contains a
      ↪  sequence of characters clearly written, distorted,
      ↪  stylized, or blended with the environment (
      ↪ background or object) as a visual illusion.
The support context only consists of top-left corner point
      ↪ of the bounding box.

**TASK**
Think and analyze the image carefully. Using the support
      ↪ context as a reference, detect the sequence of
      ↪ characters and provide a single bounding box that
      ↪ encloses all of its characters with detailed
      ↪ reasoning.

Return the result as a valid JSON list. If no sequence is
      ↪ confidently located, return empty list [].

**JSON output format**
‘‘‘json
{
    "sequence": "The sequence of characters you read and
     ↪ detected",
    "reasoning": "Explain the visual evidence for this
     ↪ detection.",
    "x1": int,
    "y1": int,
    "x2": int,
    "y2": int
}
‘‘‘
```

Prompt K: Attribution Guided Query (Celeb)

````
**INSTRUCTIONS**
You are given an image of size 1024x1024 that was generated from the above support image. A face is present in this image. The face might be clearly visible, perturbed, stylized, or blended with the environment (background or object) as a visual illusion.

**TASK**
Think and analyze the image carefully. Using the support context as a reference, detect the bounding boxes and provide detailed reasoning for each discernable facial attributes:
1. Left Eye
2. Right Eye
3. Nose
4. Mouth

Return the result as a valid JSON list. Your reasoning must explain how you identified the attribute, and only include the attribute that you can detect. If no attributes are confidently located, return empty list [].

**JSON output format**
```json
{
    "left_eye": {
        "reasoning": "Explain the visual evidence for this detection.",
        "x1": int,
        "y1": int,
        "x2": int,
        "y2": int
    },
    "right_eye": {
        "reasoning": "Explain the visual evidence for this detection.",
        "x1": int,
        "y1": int,
        "x2": int,
        "y2": int
    },
    "nose": {
        "reasoning": "Explain the visual evidence for this detection.",
        "x1": int,
        "y1": int,
        "x2": int,
        "y2": int
    },
    "mouth": {
        "reasoning": "Explain the visual evidence for this detection.",
        "x1": int,
        "y1": int,
        "x2": int,
        "y2": int
    }
}
```
````

Prompt L: Attribution Semi Guided Query (Celeb)

````
**INSTRUCTIONS**
You are given an image of size 1024x1024 that was generated from the above support image. A face is present in this image. The
    ↪ face might be clearly visible, perturbed, stylized, or blended with the environment (background or object) as a visual
    ↪ illusion.
The support context only consists of one key feature's bounding box.

**TASK**
Think and analyze the image carefully. Using the support context as a reference, detect the bounding boxes and provide detailed
    ↪ reasoning for each discernable facial attributes:
1. Left Eye
2. Right Eye
3. Nose
4. Mouth

Return the result as a valid JSON list. Your reasoning must explain how you identified the attribute, and only include the
    ↪ attribute that you can detect. If no attributes are confidently located, return empty list [].

**JSON output format**
```json
{
    "left_eye": {
        "reasoning": "Explain the visual evidence for this detection.",
        "x1": int,
        "y1": int,
        "x2": int,
        "y2": int
    },
    "right_eye": {
        "reasoning": "Explain the visual evidence for this detection.",
        "x1": int,
        "y1": int,
        "x2": int,
        "y2": int
    },
    "nose": {
        "reasoning": "Explain the visual evidence for this detection.",
        "x1": int,
        "y1": int,
        "x2": int,
        "y2": int
    },
    "mouth": {
        "reasoning": "Explain the visual evidence for this detection.",
        "x1": int,
        "y1": int,
        "x2": int,
        "y2": int
    }
}
```
````