HueManity: Probing Fine-Grained Visual Perception in MLLMs

Rynaa Grover[*1], **Jayant Sravan Tamarapalli**[*1], **Sahiti Yerramilli** [*1], **Nilay Pande**[*2]

[1]Google

[2]Waymo

{rynaa, jayantsravan, sahitiy}@google.com, nilayp@waymo.com

Abstract

Multimodal Large Language Models (MLLMs) excel at high-level visual reasoning, but their performance on nuanced perceptual tasks remains surprisingly limited. We present HueManity, a benchmark designed to assess visual perception in MLLMs. The dataset comprises 83,850 images featuring two-character alphanumeric strings embedded in Ishihara test style dot patterns, challenging models on precise pattern recognition. Our evaluation of nine state-of-the-art MLLMs on HueManity demonstrates a significant performance deficit compared to human and traditional computer vision baselines. The best-performing MLLM achieved a 33.6% accuracy on the numeric 'easy' task and a striking 3% on the alphanumeric 'hard' task. In contrast, human participants achieved near-perfect scores (100% and 95.6%)[1], and a fine-tuned ResNet50 model reached accuracies of 96.5% and 94.5%. These results highlight a critical gap in the visual capabilities of current MLLMs. Our analysis further explores potential architectural and training-paradigm factors contributing to this perceptual gap in MLLMs. We open-source HueManity dataset and code to foster further research in improving perceptual robustness of MLLMs.

Code: https://github.com/rynaa/huemanity

Dataset: https://huggingface.co/datasets/Jayant-Sravan/HueManity

1 Introduction

The trajectory of Multimodal Large Language Models (MLLMs) [Team *et al.*, 2023; Achiam *et al.*, 2023; Bai *et al.*, 2023a; Li *et al.*, 2023b; Gong *et al.*, 2023; Liu *et al.*, 2024a; Liu *et al.*, 2023; Anthropic, 2025] in recent years has been marked by impressive advancements, demonstrating sophisticated capabilities in bridging visual and textual information. Their capabilities extend well beyond simple image labeling [Russakovsky *et al.*, 2015; Deng, 2012], enabling complex tasks like generating detailed image descriptions [Dong *et al.*, 2024; Fu *et al.*, 2024a], answering intricate visual questions requiring inference about relationships and activities [Weng *et al.*, 2025; Chen *et al.*, 2025; Kuang *et al.*, 2024], and participating in nuanced dialogue about visual content [Cao *et al.*, 2024]. A key factor in their success is pre-training on vast web-scale image-text datasets, which facilitates learning powerful representations that capture high-level semantic links between visual features and language [Jia *et al.*, 2021; Radford *et al.*, 2021; Schuhmann *et al.*, 2022; Alayrac *et al.*, 2022; Qi *et al.*, 2020; Zhai *et al.*, 2022; Pham *et al.*, 2023]. Consequently, MLLMs perform strongly on tasks benefiting from this conceptual understanding, including recognizing common objects, interpreting general scene structure, and relating them to text.

However, the predominant evaluation paradigms for MLLMs have centered on these conceptual capabilities, largely overlooking their fine-grained perceptual acuity [Bai *et al.*, 2023b; Li *et al.*, 2024; Li *et al.*, 2023a; Xu *et al.*, 2024; Yin *et al.*, 2023; Liu *et al.*, 2023]. This leaves a critical gap in understanding their performance on tasks demanding precise visual discernment – such as intricate pattern recognition, subtle feature differentiation, and robust segregation within visually cluttered background. Unlike tasks solvable through broad semantic association, these perceptual challenges require a more fundamental visual processing capacity, akin to the human ability to resolve intricate details. This paper introduces a benchmark specifically designed to probe this nuanced dimension of MLLM performance. Our methodology draws inspiration from the principles of Ishihara plates [Clark, 1924], a technique traditionally employed in human ophthalmology to assess color vision by embedding figures (like numbers or paths) within fields of multicolored, varied-size dots. It is crucial to clarify that HueManity does not aim to diagnose 'color blindness' in MLLMs. Instead, our Ishihara-style stimuli, created using controlled generation techniques, rigorously test an MLLM's fundamental ability to identify embedded alphanumeric characters by their subtle color and luminance contrasts within visually cluttered dot patterns.

Success on the HueManity benchmark serves as a crucial indicator of an MLLM's potential for robust visual understanding in complex, real-world scenarios. Unlike often cu-

[*]Equal Contribution.

[1]Human evaluations utilized 100-image representative subsets, sampled from the model evaluation sets for each respective task.

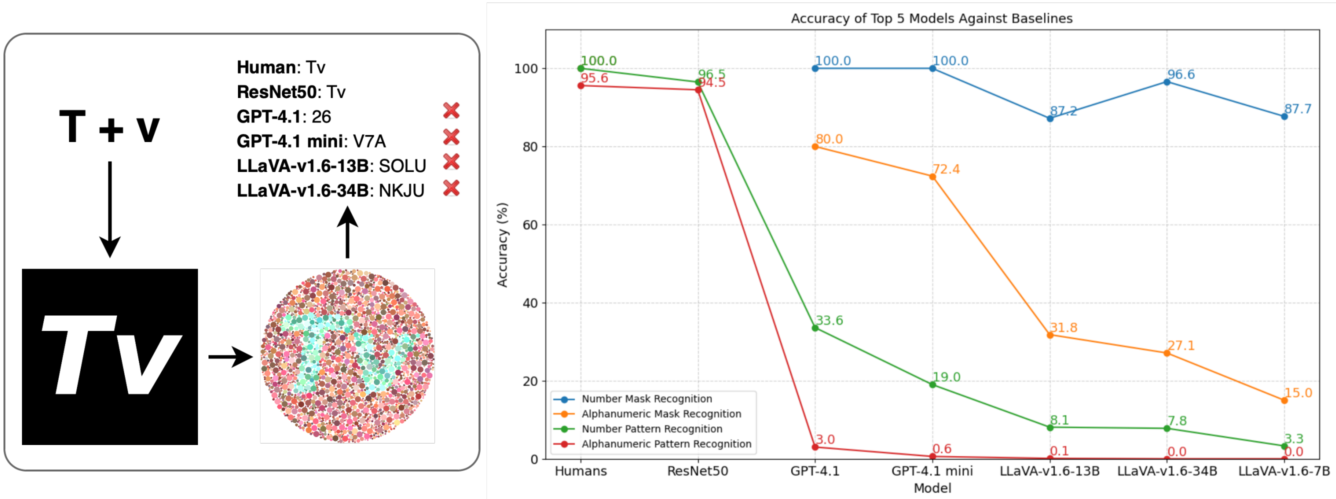


Figure 1: **HueManity** — A new benchmark for MLLM fine-grained visual perception. The pipeline creates a character mask from alphanumeric characters and renders it as an Ishihara-style pattern. While models achieve high accuracy on clear masks, they struggle on the challenging pattern images.

rated benchmark datasets, real-world visual environments are frequently characterized by clutter, partial occlusions, variable lighting, and unconventional information presentation. The ability of an MLLM to reliably parse characters in our Ishihara-style plates is intended to assess its resilience to visual clutter and its pattern recognition capabilities — foundational skills often linked to dependable performance in challenging visual settings. Furthermore, mastering this level of fine-grained perception is foundational for more intricate forms of visual reasoning. An MLLM's struggles with basic pattern recognition under challenging conditions could imply difficulties in performing higher-order tasks that depend on the accurate interpretation of subtle details. Thus, HueManity serves not merely as a test of pattern recognition, but as a probe into architectural or training limitations that hinder MLLMs from achieving comprehensive, detailed visual intelligence.

To address this identified gap and facilitate further research in this domain, this paper makes the following specific contributions:

1. **We introduce HueManity, a new large-scale benchmark (83,850 images)** featuring Ishihara-inspired alphanumeric stimuli. The benchmark utilizes a principled design with 25 carefully curated color pairs, selected using CIEDE2000 (ΔE_{2000}) metrics and manual verification, ensuring both systematic challenge and fairness for human comparison.
2. **We conducted a comprehensive evaluation of nine state-of-the-art MLLMs on HueManity's easy and hard sets.** Findings reveal a significant performance gap compared to strong human and fine-tuned ResNet50 baselines. Baseline success suggests MLLM limitations are due to their current architectures or training for fine-grained perception, rather than the task being intractable.
3. **We release open-source code for generating challenging Ishihara-style perceptual stimuli** (alphanumeric characters embedded in dot patterns), enabling reproducible research and community-driven extensions.

2 Related Works

2.1 Multimodal Models

With the remarkable advancements of Large Language Models (LLMs), recent research has extended their capabilities to multimodal domains by integrating visual information, giving rise to Multimodal Large Language Models (MLLMs) [Team *et al.*, 2023; Achiam *et al.*, 2023; Bai *et al.*, 2023a; Li *et al.*, 2023b; Gong *et al.*, 2023; Liu *et al.*, 2024a; Liu *et al.*, 2023]. These models typically align visual features from pre-trained image encoders with LLMs via modality adaptation layers. Early works like BLIP-2 [Li *et al.*, 2023b] pioneered this architecture by first pre-training on image-text datasets and fine-tuning on task-specific benchmarks such as Visual Question Answering (VQA). Subsequent models like LLaVA [Liu *et al.*, 2023] advanced this approach by leveraging synthetic instruction-following data in VQA formats, significantly improving instruction tuning performance. More recent efforts have expanded into video understanding and even image generation, showcasing the versatility of MLLMs across modalities. However, much of their celebrated success in visual tasks often appears intertwined with, and perhaps reliant upon, their powerful language capabilities to interpret, reason about, and generate text from visual content. This inherent strength in linguistic processing may have led to less emphasis on developing or scrutinizing their more fundamental, non-linguistic visual perception skills. HueManity directly addresses this gap in scrutinizing MLLM visual perception. It is a benchmark specifically designed to isolate and rigorously evaluate these core abilities.

2.2 MLLM Evaluation

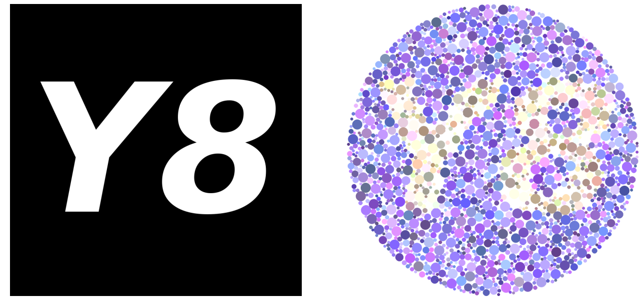
Multimodal Large Language Models (MLLMs) are highly capable of global image understanding and reasoning but consistently struggle with fine-grained visual tasks such as precise recognition and localization [Huang and Zhang, 2024; Li *et al.*, 2024]. Several benchmarks have been proposed to address these challenges. TouchStone [Bai *et al.*, 2023b] offers 908 manually annotated visual dialog questions across five abilities and 27 sub-tasks, while LLaVA-Bench [Liu *et al.*, 2023] includes 24 images with 60 curated questions covering diverse content like scenes, memes, and sketches. To automate evaluation, LLaVA-Bench [Liu *et al.*, 2023], LAMM [Yin *et al.*, 2023], and TouchStone [Bai *et al.*, 2023b] rely on GPT-based models to judge relevance and accuracy, but this introduces inherent reliability issues and cost inefficiencies. Similarly, LVLM-eHub [Xu *et al.*, 2024] aggregates multiple vision benchmarks but still depends on human annotators to compare model outputs, making it both subjective and expensive. More structured efforts like MME and MMBench [Liu *et al.*, 2024b] attempt to provide objective evaluations with multiple-choice questions across a wide range of ability dimensions. MME pioneered multimodal Yes/No questions for perception and reasoning, with MM-Vet [Yu *et al.*, 2023] and MMBench [Liu *et al.*, 2024b] extending coverage to sub-tasks like OCR, math, and recognition, yet they all heavily rely on existing VQA datasets or GPT-generated questions. SEED-Bench [Li *et al.*, 2024; Li *et al.*, 2023a] scales up to 24,000 human-annotated multiple-choice questions covering varied input-output modalities, but its challenges remain simple, with most open-source models reaching 30–60% accuracy at the easiest level. Blink [Fu *et al.*, 2024b] attempts a more holistic evaluation across 14 perception tasks with 3,900 questions and 7,300 images, but does not assess models on combinations of tasks, and remains less challenging, as evidenced by higher model accuracies.

HueManity uniquely evaluates Multimodal Large Language Models (MLLMs) on the foundational visual skill of discerning patterns from subtle cues in cluttered environments, employing alphanumeric characters within Ishihara-style dot patterns with controlled color contrasts as a robust proxy. This stimulus design, combined with procedural generation and exact-match evaluation, establishes a scalable, objective, and reliable methodology distinct from more subjective or resource-intensive techniques.

3 Data Creation

The HueManity dataset comprises 83,850 images, each pairing a two-character alphanumeric string with its ground truth label and full generation parameters. These strings are formed from lowercase letters (a-z), uppercase letters (A-Z), and digits (0-9). To enhance clarity and prevent evaluation errors, we excluded visually ambiguous characters (‘l’, ‘I’, ‘J’, ‘O’) and combinations commencing with ‘0’ (e.g., ‘01’, to avoid ‘1’ vs. ‘01’ prediction conflicts). The entire dataset was produced by rendering all valid two-character combinations using 25 meticulously curated color pairs, which then formed the basis for our evaluation sets.

The full HueManity dataset, encompassing all 83,850 generated images, serves as a comprehensive resource for evaluating MLLM perception. However, for the specific MLLM evaluations reported in this paper, particularly considering practical constraints such as time and API access costs, we defined two distinct tasks using carefully sampled subsets. Each subset comprises 1,000 images, randomly selected to be representative for these focused evaluations:



(a) Binary Text Mask (b) Ishihara pattern

Figure 2: Image generation process for the alphanumeric string “Y8”. (a) The binary text mask. (b) The corresponding Ishihara-style dot pattern.

1. **Number Recognition Set (Used for the ‘Easier’ Task):** This subset of 1,000 images is used to evaluate models on the **Number Identification task**. It exclusively contains images displaying two-character strings composed solely of numeric digits (*e.g.*, ‘17’, ‘83’, ‘05’).
2. **Text Recognition Set (Used for the ‘Harder’ Task):** This subset of 1,000 images, drawn from the broader pool of alphanumeric combinations, is employed for the **Text Identification task**. It features diverse two-character alphanumeric strings, including various combinations of letters and digits (*e.g.*, ‘A7’, ‘b9’, ‘XG’).

3.1 Text Mask Generator

The initial stage in our data pipeline is the creation of a 900x900 pixel resolution binary text mask for each two-character string. Using Pygame, we render these strings in white against a black background. To ensure both clear readability and adequate character width suitable for the subsequent dot-based rendering, all masks utilize the DejaVu Sans font (size 550, bold and italic styles), as exemplified in Figure 2a.

3.2 Ishihara-Style Pattern Generation

Our pattern generator, adapted from an open-source Pygame project[2], iteratively populates the image with non-overlapping circles. Over 30,000 iterations, the generator performs the following steps for each potential circle:

1. **Placement:** Randomly sample (x,y) coordinates and compute the maximum non-colliding radius (ranging from 4 to 15 pixels).

[2]https://github.com/hakrackete/Ishihara-color-plate-generator

2. **Initial Color:** Assign initial color (foreground/background) based on whether the center falls within the character mask (Fig. 2a).
3. **Color Transformation:** Apply three randomized transformations to the initial color: gradient shift towards the other pair color, RGB color shift (range $[-30, +30]$), and RGB lightness scaling (factor 0.66-1.5).
4. **Rendering:** Render the transformed circle at the computed position.

This iterative process ultimately yields the final, dense Ishihara-style pattern (as shown in Figure 2b).

3.3 Color Pairs Selection

The 25 distinct foreground-background color pairs used in the HueManity stimuli were meticulously selected through a multi-stage process to ensure a balance between perceptual challenge and unambiguous human legibility. This procedure involved quantitative CIEDE2000 [Luo *et al.*, 2001] analysis and extensive manual verification, with the full methodology detailed in Appendix B.

4 Experiments

We evaluated a diverse set of nine state-of-the-art Multimodal Large Language Models (MLLMs), encompassing prominent commercial APIs and publicly available open-source models. Model inference was orchestrated using Promptfoo[3], a platform facilitating reproducible benchmarking through flexible prompt definition and API integration. Open-source models were hosted locally via Ollama[4] and inferred on a single NVIDIA A100 GPU.

Evaluation focused on two primary tasks: numerical recognition (digits only) and alphanumeric recognition (digits and letters). For each of these tasks, a subset of 1,000 points from HueManity: two-character strings and their corresponding generated images were used. Models were evaluated on two types of visual stimuli for each datapoint:

- **HueManity Plates:** The 1,000 randomly sampled Ishihara-style dot pattern images from the HueManity benchmark.
- **Text Masks:** The corresponding 1,000 binary text mask images (i.e., white text on a black background, as exemplified in Figure 2a). This additional evaluation on text masks was performed to establish a baseline for each model's ability to recognize the characters in a clear, unobstructed format, helping to differentiate fundamental OCR capabilities from performance on the perceptually challenging dot patterns.

All images (both HueManity plates and text masks) were provided to the models encoded in Base64 format. The following task-specific prompts were used for both stimulus types:

- **Number Recognition Prompt:** "What is the number in this image? Strictly stick to the format: Answer: [number in the image]"
- **Text Recognition Prompt:** "What is the exact text in this image? It has only alpha-numeric characters excluding small L, capital O, capital I, and capital J to avoid ambiguity. Strictly stick to the format: Answer: [exact text in the image]"

[3]https://www.promptfoo.dev

[4]https://ollama.com

4.1 Human Performance Evaluations

To establish a human performance baseline, evaluations involved three adult volunteers, all with self-reported normal color vision. This group size was considered adequate because identifying characters in HueManity plates is a rudimentary perceptual task for humans with normal vision, one that requires no specialized skills and typically exhibits low inter-volunteer performance variance. For each of the two tasks (numerical and alphanumeric recognition), these volunteers were tested on a representative subset of 100 images, randomly sampled from the corresponding 1,000-image evaluation sets used for MLLMs.

Volunteers viewed the images at their original 900x900 pixel resolution within a Google Sheets document. Using the identical prompts provided to the MLLMs, they were asked to identify the embedded characters, entering their responses directly into the document. This methodology ensured a direct comparison of human and MLLM performance under equivalent conditions.

4.2 Traditional Computer Vision Baseline (ResNet50)

As a representative traditional computer vision baseline, we trained and evaluated a ResNet50 model. We utilized a ResNet50 pre-trained on ImageNet, obtained from the PyTorch `vision` library.

The standard classification layer of the ResNet50 was replaced with two independent classification heads. Each head was designed to predict one character, treating the task as two independent character recognition problems. For the purpose of fine-tuning this model, we utilized 2,000 images randomly sampled from the broader HueManity dataset, ensuring these were distinct from the final evaluation subsets. Training was conducted for 30 epochs using the Adam optimizer with a learning rate of $1e-3$. The loss function for training was the sum of the cross-entropy losses calculated independently for each of the two classification heads. The trained model was evaluated on the same 1,000-image subsets used for the MLLM evaluations.

5 Results and Analysis

5.1 Human Performance: A Near-Perfect Baseline

Human volunteers established a crucial performance baseline, demonstrating exceptionally high accuracy and efficiency on the HueManity benchmark (detailed in Table 3). On the numerical recognition task, volunteers achieved a perfect **100%** average accuracy, indicating that identifying two-digit strings within the dot patterns was straightforward. The more complex alphanumeric task also yielded a very strong average accuracy of **95.6%**.

The minor errors in the alphanumeric task were insightful: annotator feedback consistently attributed confusion not to an inability to perceive the characters, but to difficulty differentiating **visually similar character forms** (*e.g.*, 's', 'c', 'w' upper and lower case variants). This distinction is important, as it underscores that the primary perceptual challenge of extracting patterns from the dots was readily overcome by the human visual system. Furthermore, volunteers processed these images with remarkable efficiency, typically requiring **less than one second per image** to identify the characters.

Near-perfect and rapid human scores offer crucial insights into MLLM performance on this task for several key reasons:

- **Confirms Task Solvability and Stimulus Clarity:** High human accuracy confirms the stimuli provide unambiguous visual information for character identification, showing the tasks are well-defined and solvable by proficient visual systems.
- **Establishes a Performance Ceiling:** Human performance sets a clear accuracy ceiling, highlighting the target for machine perception systems on this (for humans) perceptually simple task.
- **Highlights the Nature of MLLM Deficiencies:** Significant MLLM deviations from human scores indicate machine-specific limitations, not task impossibility or unclear stimuli. This contrast helps determine if MLLM struggles lie in initial perceptual grouping or later recognition.
- **Contextualizes Machine Errors:** Human ease and minor confusions (*e.g.*, similar glyphs) provide valuable context for analyzing MLLM error types and severity, pinpointing where machine perception diverges from human visual understanding.

In essence, the human baseline shows HueManity tests foundational visual skills that humans execute efficiently, despite MLLM challenges. This context is vital for appreciating the scale and nature of MLLM performance gaps.

5.2 ResNet50 Baseline: Demonstrating Task Learnability

To further contextualize the performance of Multimodal Large Language Models (MLLMs), we established a strong baseline using a fine-tuned ResNet50 [He *et al.*, 2015], a well-established traditional computer vision architecture. As detailed in Table 1 and Table 2, the ResNet50 model achieved high accuracy, scoring **96.5% on the numerical recognition task and 94.5% on the alphanumeric task**.

This robust performance from a standard convolutional architecture, fine-tuned on a relatively small set of 2,000 images from the HueManity dataset (as described in Section 4.2), is highly informative. It clearly demonstrates that the task of identifying characters within the HueManity dot patterns is fundamentally learnable by established computer vision techniques.

Furthermore, achieving near-human accuracy with this setup suggests that the perceptual cues embedded in the stimuli are sufficiently rich and consistent for a focused model to learn effective recognition. This implies that the task, while designed to challenge nuanced perception, is **not inherently intractable for AI and can be considered relatively easy for appropriately adapted vision models**. The ResNet50 baseline's success, therefore, pinpoints the challenges as MLLM-specific. This suggests that MLLM difficulties likely arise not from an inherently unsolvable visual task, but from how these larger, more general models currently process or prioritize fine-grained perceptual information.

	Mask	**Patterned Image**
Humans (average)	100%	100%
ResNet50	-	96.5%
API-based models		
GPT-4.1 mini	100%	19.0%
GPT-4.1	100%	33.6%
Claude 3.7 Sonnet	100%	0.4%
Open-source models		
LLaVA-v1.6-7B	87.7%	3.3%
LLaVA-v1.6-13B	87.2%	8.1%
LLaVA-v1.6-34B	96.6%	7.8%
Mistral-small3.1-24b	100%	0.1%
Qwen VL Max	100%	0.2%
Pixtral	100%	1%

Table 1: **Accuracy on the number recognition task** for human evaluators, ResNet50, and various MLLMs on both text masks and patterned HueManity images.

	Mask	**Patterned Image**
Humans (average)	100%	95.6%
ResNet50	-	94.5%
API-based models		
GPT-4.1 mini	72.4%	0.6%
GPT-4.1	80%	3.0%
Claude 3.7 Sonnet	82.2%	0%
Open-source models		
LLaVA-v1.6-7B	15%	0%
LLaVA-v1.6-13B	31.8%	0.1%
LLaVA-v1.6-34B	27.1%	0%
Mistral-small3.1-24b	58.7%	0%
Qwen VL Max	83.5%	0%
Pixtral	65.8%	1.8%

Table 2: **Accuracy on the alphanumeric recognition task**, comparing human performance with ResNet50 and MLLMs across text masks and patterned HueManity images.

5.3 The Perceptual Gap: MLLM Performance vs. Baselines on HueManity

The performance of the nine evaluated Multimodal Large Language Models (MLLMs) on the HueManity benchmark stands in stark contrast to the near-perfect accura-

	Number Pattern	Alphanumeric Pattern
Evaluator 1	100%	97%
Evaluator 2	100%	95%
Evaluator 3	100%	95%
Average	100%	95.6%

Table 3: Human evaluation accuracies on 100-image subsets for both number and alphanumeric pattern recognition tasks.

cies achieved by both human volunteers and the fine-tuned ResNet50 baseline [He *et al.*, 2015] (detailed in Table 1 and Table 2). Across both the numerical and alphanumeric tasks, all evaluated MLLMs exhibited a profound and consistent inability to match these baselines. For instance, even the best-performing MLLM achieved only **33.6% on the easier numeric task and a mere 3% on the harder alphanumeric task**—figures that dramatically underscore their struggle.

This significant and consistent underperformance, on tasks demonstrably solvable by human vision and traditional AI methodologies, signals a critical gap in the fine-grained visual perception capabilities of current MLLMs. Such results necessitate a deeper exploration into the potential underlying reasons for these models' difficulties.

Several characteristics inherent to the current design and training paradigms of many Multimodal Large Language Models (MLLMs) may contribute to their observed difficulties on tasks demanding nuanced visual perception. Firstly, the **visual processing pipeline itself could be a factor**. Vision encoders in MLLMs, though often powerful and pre-trained on diverse datasets (*e.g.*, ViT variants [Liu *et al.*, 2023; Liu *et al.*, 2024a; Agrawal *et al.*, 2024; Bai *et al.*, 2023a; Bai *et al.*, 2025], are typically optimized for capturing semantic information and global scene context. This optimization focus might lead to the unintentional down-weighting or loss of extremely fine-grained local details—like subtle color shifts defining patterns within a dense, similarly featured background—that are crucial when strong semantic cues are absent. Furthermore, the interface used to connect these vision encoders to the language model—often a projection layer or a small MLP [Liu *et al.*, 2023; Liu *et al.*, 2024a; Agrawal *et al.*, 2024; Bai *et al.*, 2023a; Bai *et al.*, 2025]—necessarily transforms and condenses visual information. This critical step, while enabling multimodal fusion, **might inadvertently act as an information bottleneck, potentially abstracting or losing the precise, high-resolution feature distinctions** necessary for meticulous perceptual tasks.

Secondly, the **nature of MLLM pre-training data and their resulting operational strengths may not fully align with the demands of certain fundamental perception challenges.** MLLMs are typically pre-trained on vast web-scale corpora where images are often paired with textual descriptions or interleaved image-text documents [Liu *et al.*, 2023; Liu *et al.*, 2024a; Agrawal *et al.*, 2024; Bai *et al.*, 2023a; Bai *et al.*, 2025; Achiam *et al.*, 2023]. Such datasets, while fostering robust semantic alignment and impressive contextual understanding, might underrepresent stimuli requiring intensive perceptual organization purely from low-level visual features, without strong linguistic or clear object-based anchors. Consequently, MLLMs may not have adequately developed the specialized visual routines needed for tasks like consistently grouping spatially distributed elements based only on subtle shared properties (*e.g.*, color similarity) within a field of similar distractors. Their celebrated success in many multimodal tasks often stems from leveraging their powerful integrated Large Language Models (LLMs) to interpret, reason about, and generate language from visual input [Achiam *et al.*, 2023; Anthropic, 2025; Liu *et al.*, 2023; Gong *et al.*, 2023; Jiang *et al.*, 2023]. This reliance on higher-level, often text-mediated, understanding can be less effective when the core challenge demands direct, bottom-up visual processing and pattern extraction, rather than conceptual inference or semantic association. **Such tasks require a foundational visual acuity that may not be a primary emergent outcome of training regimes largely focused on multimodal semantics and instruction following.**

6 Conclusion and Future Directions

In this work, we introduced HueManity, a new benchmark dataset comprising 83,850 Ishihara-style images designed to rigorously evaluate the fine-grained visual perception of Multimodal Large Language Models (MLLMs). Our comprehensive evaluations of nine MLLMs revealed a stark performance gap: these advanced models struggled significantly (e.g., 3-34% accuracy) on tasks readily solved by human volunteers (95-100% accuracy) and traditional computer vision models like a fine-tuned ResNet50 (94-96% accuracy). This highlights a critical limitation in current MLLM capabilities to discern patterns based on subtle visual cues in complex, noisy environments. We provide the dataset and generation code to the community to foster further research.

The observed MLLM deficiencies likely stem from a combination of factors, including vision encoders not optimized for such fine-grained detail, information bottlenecks at the vision-language interface, pre-training data that underrepresents such perceptual challenges, and an over-reliance on text-mediated or high-level semantic reasoning for tasks demanding direct, bottom-up visual processing. Future work should therefore focus on several key directions to bridge this perceptual gap: 1) Developing novel MLLM architectures with improved visual front-ends and fusion mechanisms that better preserve and process low-level visual information. 2) Augmenting pre-training and fine-tuning datasets with stimuli specifically designed to enhance robust perception in visually complex scenarios. 3) Exploring training objectives that explicitly encourage the development of fundamental visual acuity and perceptual organization skills, independent of, yet complementary to, high-level reasoning. Addressing these areas will be crucial for advancing MLLMs towards more holistic and human-like visual understanding.

7 Limitations

While HueManity provides valuable insights into a specific facet of MLLM visual perception, we acknowledge certain limitations. Firstly, the benchmark evaluates a focused perceptual task: identifying two-character alphanumeric strings within Ishihara-style dot patterns. Although this serves as a robust proxy for the ability to discern patterns from visually complex backgrounds based on subtle cues, its direct generalization to the full spectrum of diverse, real-world fine-grained perceptual challenges MLLMs may encounter requires further investigation. Secondly, our current dataset, while substantial and systematically generated with controlled color contrasts, primarily explores variations in color and basic character forms. Future iterations or complementary benchmarks could expand to assess a wider array of perceptual dimensions, such as sensitivity to texture, orientation, motion, or more intricate compositions of embedded elements. Finally, while our study includes a diverse set of nine contemporary MLLMs, the MLLM landscape is evolving rapidly. Consequently, our findings represent a snapshot based on the models and their versions tested at the time of this study, and newer or differently architected models might exhibit different performance characteristics.

8 Usage of Generative AI tools

We utilized Generative AI tools to help improve the language, phrasing, and readability of this manuscript.

References

[Achiam *et al.*, 2023] 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.

[Agrawal *et al.*, 2024] Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, Timothée Lacroix, Guillaume Lample, Diego Las Casas, Thibaut Lavril, Teven Le Scao, Andy Lo, William Marshall, Louis Martin, Arthur Mensch, Pavankumar Muddireddy, Valera Nemychnikova, Marie Pellat, Patrick Von Platen, Nikhil Raghuraman, Baptiste Rozière, Alexandre Sablayrolles, Lucile Saulnier, Romain Sauvestre, Wendy Shang, Roman Soletskyi, Lawrence Stewart, Pierre Stock, Joachim Studnia, Sandeep Subramanian, Sagar Vaze, Thomas Wang, and Sophia Yang. Pixtral 12b, 2024.

[Alayrac *et al.*, 2022] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022.

[Anthropic, 2025] Anthropic. Claude 3.5 sonnet, 2025. Accessed May 18, 2025.

[Bai *et al.*, 2023a] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023.

[Bai *et al.*, 2023b] Shuai Bai, Shusheng Yang, Jinze Bai, Peng Wang, Xingxuan Zhang, Junyang Lin, Xinggang Wang, Chang Zhou, and Jingren Zhou. Touchstone: Evaluating vision-language models by language models. *arXiv preprint arXiv:2308.16890*, 2023.

[Bai *et al.*, 2025] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025.

[Cao *et al.*, 2024] Qingxing Cao, Junhao Cheng, Xiaodan Liang, and Liang Lin. VisDiaHalBench: A visual dialogue benchmark for diagnosing hallucination in large vision-language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12161–12176, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[Chen *et al.*, 2025] Boqi Chen, Anuj Khare, Gaurav Kumar, Arjun Akula, and Pradyumna Narayana. Seeing beyond: Enhancing visual question answering with multi-modal retrieval. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, Steven Schockaert, Kareem Darwish, and Apoorv Agarwal, editors, *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 410–421, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.

[Clark, 1924] JH Clark. The ishihara test for color blindness. *American Journal of Physiological Optics*, 1924.

[Deng, 2012] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[Dong *et al.*, 2024] Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, Xiangwen Kong, Xiangyu Zhang, Kaisheng Ma, and Li Yi. Dreamllm: Synergistic multimodal comprehension and creation, 2024.

[Fu *et al.*, 2024a] Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. Guiding instruction-based image editing via multimodal large language models, 2024.

[Fu *et al.*, 2024b] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive.

In *European Conference on Computer Vision*, pages 148–166. Springer, 2024.

[Gong *et al.*, 2023] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*, 2023.

[He *et al.*, 2015] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[Huang and Zhang, 2024] Jiaxing Huang and Jingyi Zhang. A survey on evaluation of multimodal large language models. *arXiv preprint arXiv:2408.15769*, 2024.

[Jia *et al.*, 2021] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision, 2021.

[Jiang *et al.*, 2023] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

[Kuang *et al.*, 2024] Jiayi Kuang, Jingyou Xie, Haohao Luo, Ronghao Li, Zhe Xu, Xianfeng Cheng, Yinghui Li, Xika Lin, and Ying Shen. Natural language understanding and inference with mllm in visual question answering: A survey, 2024.

[Li *et al.*, 2023a] 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.

[Li *et al.*, 2023b] 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.

[Li *et al.*, 2024] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308, 2024.

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

[Liu *et al.*, 2024a] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.

[Liu *et al.*, 2024b] 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.

[Luo *et al.*, 2001] Ming Luo, Guihua Cui, and B. Rigg. The development of the cie 2000 colour-difference formula: Ciede2000. *Color Research Application*, 26:340 – 350, 10 2001.

[Pham *et al.*, 2023] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, Mingxing Tan, and Quoc V. Le. Combined scaling for zero-shot transfer learning, 2023.

[Qi *et al.*, 2020] Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data, 2020.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[Russakovsky *et al.*, 2015] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge, 2015.

[Schuhmann *et al.*, 2022] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022.

[Team *et al.*, 2023] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[Weng *et al.*, 2025] Weixi Weng, Jieming Zhu, Xiaojun Meng, Hao Zhang, Rui Zhang, and Chun Yuan. Learning to compress contexts for efficient knowledge-based visual question answering, 2025.

[Xu *et al.*, 2024] Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[Yin *et al.*, 2023] Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Xiaoshui Huang, Zhiyong Wang, Lu Sheng, Lei Bai, et al. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *Advances in Neural Information Processing Systems*, 36:26650–26685, 2023.

[Yu *et al.*, 2023] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

[Zhai *et al.*, 2022] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning, 2022.

A A Brief Discussion on CIEDE2000 Color Difference

The CIEDE2000 score (ΔE_{2000}, Equation 1) [Luo *et al.*, 2001] quantifies the perceived difference between two colors more accurately than prior formulae, especially for subtle variations. It calculates a single value representing the "distance" between colors in the perceptually uniform CIE $L^*a^*b^*$ space, considering lightness, chroma, and hue. In the HueManity benchmark, ΔE_{2000} was pivotal for systematically designing stimuli. The ability to discern characters in the Ishihara-style plates directly depends on the perceived color contrast between foreground (character) and background dots. This score provided a perceptually relevant, objective method to quantify this contrast, enabling the selection of color pairs across a controlled spectrum of difficulty, refer to Figure 3. This ensured the benchmark could rigorously test visual perception for varying degrees of color discriminability while maintaining stimuli legibility for human comparison, forming a foundational aspect of our dataset's controlled experimental design.

B Color Pairs Selection

The selection of appropriate color pairs for the foreground (characters) and background dots was a critical phase in the development of HueManity, undertaken with considerable care to ensure a balance between perceptual challenge and unambiguous human legibility. The process involved several stages:

1. **Initial Candidate Generation:** We bootstrapped the process with 15 medium-contrast color pairs generated by LLMs (Gemini, ChatGPT). This initial pool was iteratively refined by evaluating pairs against CIEDE2000 (ΔE_{2000}, Eq. 1) scores and visual checks. We retained promising candidates, modified some, and discarded others, while simultaneously manually crafting and vetting new pairs to meet the benchmark's final requirements (detailed below). This refinement cycle culminated in the selection of 25 distinct pairs for the subsequent validation stages.

2. **Quantitative Contrast Filtering (CIEDE2000):** Each of these candidate pairs then underwent rigorous quantitative analysis using the CIEDE2000 (ΔE_{2000}) color difference formula (Equation 1). This formula is a standard measure in color science, designed to reflect perceptually meaningful differences as perceived by humans. We established a specific target range for the ΔE_{2000} score, retaining only pairs with contrast values between 25 and 75. The lower bound of 25 was set to ensure sufficient theoretical distinguishability for individuals with normal color vision, preventing pairs that would be inherently ambiguous. The upper bound of 75 aimed to exclude pairs with excessively high contrast, which might render the perceptual task trivial and deviate from the subtle challenge intended.

3. **Balanced Contrast Distribution:** A key objective during selection was to ensure the benchmark included stimuli across a spectrum of difficulty levels related to color similarity. Therefore, we deliberately curated the final set of 25 color pairs to achieve an approximately equal distribution around a ΔE_{2000} score of 50. This threshold is grounded in color science literature, often considered a point distinguishing more subtle (scores $<$50) from more clearly distinct (scores $>$50) color differences. We aimed for roughly half the selected pairs to fall below this threshold and half above, ensuring HueManity evaluates performance across varying, literature-informed degrees of color contrast difficulty.

4. **Manual Verification and Legibility Check:** Recognizing that a single numerical contrast score like ΔE_{2000} captures overall perceived difference but may not fully account for the complex interplay of hue, saturation, and luminance components, especially when rendered as dots and subjected to further transformations (gradient, color, and light shifts as described in Section 3.2), a crucial final step of manual verification was performed. As you noted, it is hard to quantify the nuanced visual impact of these combined factors with a single metric. Therefore, for every color pair that passed the quantitative filtering, sample HueManity images were generated. These renderings were meticulously inspected by the authors. The primary goal was to reject pairs where the characters, despite an acceptable overall contrast score, appeared visually too similar to the background due to the specific combination of hue, saturation, luminance, or the effect of the applied shifts. This ensured that the embedded alphanumeric characters were clearly legible and that the pattern recognition was unambiguous for human observers with normal color vision. Any pairs that resulted in ambiguous characters or were otherwise problematic during this visual check were discarded.

This multi-stage process, combining LLM-based idea generation, principled quantitative filtering based on color science, a balanced distributional strategy, and crucial human judgment to account for complex visual interactions, resulted in the final curated set of 25 color pairs. This ensures that the stimuli used in HueManity are not only theoretically sound but also practically validated for fairness, legibility, and the intended level of perceptual challenge.

$$\Delta E_{2000} = \sqrt{\left(\frac{\Delta L'}{K_L S_L}\right)^2 + \left(\frac{\Delta C'}{K_C S_C}\right)^2 + \left(\frac{\Delta H'}{K_H S_H}\right)^2 + R_T \left(\frac{\Delta C'}{K_C S_C}\right)\left(\frac{\Delta H'}{K_H S_H}\right)}$$

$$\begin{aligned} \text{where} \quad & \Delta L' \text{ is the corrected lightness difference,} \\ & \Delta C' \text{ is the corrected chroma difference,} \\ & \Delta H' \text{ is the corrected hue difference,} \\ & K_L, K_C, K_H \text{ are parametric factors (typically 1),} \\ & S_L, S_C, S_H \text{ are weighting functions for lightness, chroma, and hue,} \\ & R_T \text{ is a rotation term accounting for hue-chroma interaction.} \end{aligned} \tag{1}$$

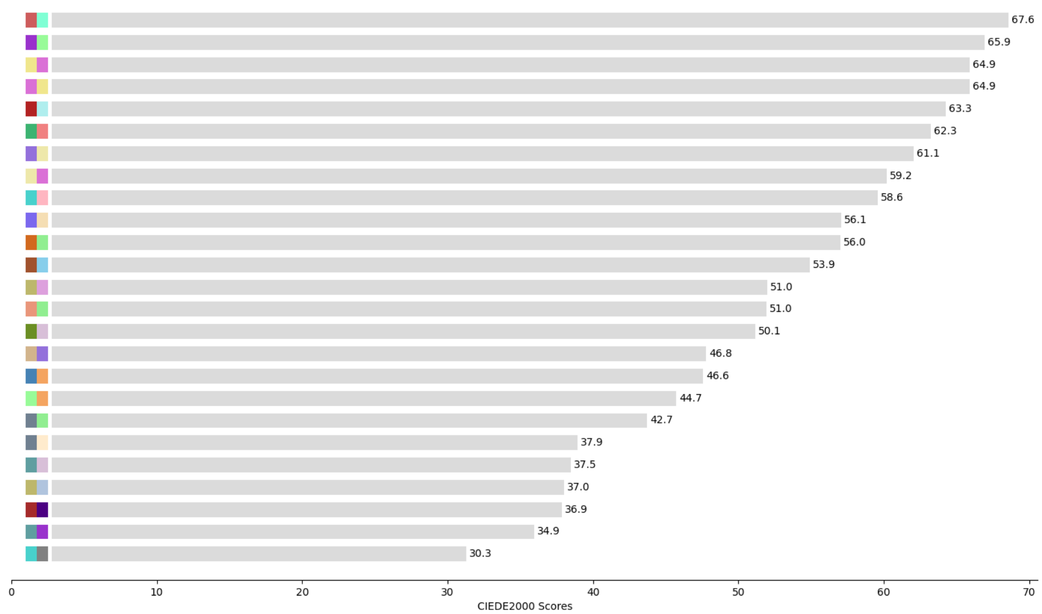


Figure 3: Distribution of CIEDE2000 color difference scores for the 25 selected foreground-background color pairs utilized in the HueManity benchmark.

C Qualitative Analysis of MLLM Failure Patterns

This section details common failure patterns observed in Multimodal Large Language Models (MLLMs) when tasked with identifying alphanumeric characters embedded in Ishihara-style dot patterns from the HueManity dataset. These observations stem from a comparative analysis of MLLM responses against human performance and ground truth data. Notably, human visual perception proved highly accurate on these tasks, with any infrequent errors typically involving confusion between graphically similar characters. In contrast, MLLMs exhibited distinct and more fundamental failure modes.

C.1 Prevalent Hallucination of Unrelated or Overly Complex Characters

A dominant failure mode across multiple MLLMs was the generation of characters, words, or even entire phrases that bore no resemblance to the two-character ground truth. This phenomenon of "hallucination" often resulted in outputs significantly more complex or contextually incongruous than the target stimuli. For instance, in the alphanumeric task, a model such as `Claude 3.7 Sonnet` might interpret a simple two-letter combination as a short phrase (e.g., responding with "MUST SEE" or "SOLU" for simple targets like "Rw" or "Tv"). Similarly, `llava-7b` could produce non-sensical strings like "HQJHSTOS", and `LLaVA-13b` occasionally generated contextually unrelated phrases like "[G3T1NGST4RT3D]". The numeric task was not immune — for a two-digit number, `Claude 3.7 Sonnet` was observed to list a sequence of unrelated two-digit numbers. This pattern suggests that when the fine-grained perceptual challenge overwhelms the MLLMs' visual processing, they may default to generating text that, while perhaps linguistically plausible, is detached from the actual visual content.

C.2 Frequent Resort to Descriptive Evasion or Explicit Admission of Inability

Rather than consistently attempting to identify the embedded characters, many MLLMs frequently defaulted to one of two evasive strategies: providing a general description of the image (often correctly identifying it as a color vision test) or explicitly stating their incapacity to discern any characters. This behavior contrasted significantly with human participants, who invariably attempted the identification task. For example, models like `GPT-4.1 Mini` and `Mistral-small3.1-24b`, when presented with alphanumeric stimuli, often responded by describing the image as an Ishihara test but stated they could not clearly identify specific characters. In the numeric task, `Claude 3.7 Sonnet` sometimes offered similar descriptive evasions, asserting no number was visible and describing the circular dot pattern. Furthermore, some models, such as `LLaVA-34b`, occasionally provided categorical statements of inability, indicating they could not recognize or interpret images and requesting a description or textual input instead. This pattern suggests that MLLMs may possess internal confidence thresholds that, when triggered by low-confidence visual parsing, lead to evasive or pre-programmed "unable to process" responses rather than a forced, best-guess attempt at character recognition.

C.3 Erratic, Unpredictable, and Systematically Flawed Output Patterns

MLLM outputs were frequently characterized by their erratic and unpredictable nature. This included the generation of seemingly random strings of characters, peculiar systematic but incorrect patterns, or extreme numerical inventions far removed from the two-character target. This high variance in error types was observed both across different models for the same input and within the outputs of a single model across different images. For instance, when presented with the same alphanumeric target (e.g., "Wh"), while one model (`GPT-4.1`) might respond almost correctly, others exhibited diverse failures: `Claude 3.7 Sonnet` produced an unrelated number ("4726"); `LLaVA-13b` generated an exceptionally long string of sequential numbers; and `Qwen VL Max` incorrectly reasoned the presence of a different number ("12"). Some incorrect outputs also suggested flawed systematic processing, such as `LLaVA-13b` responding with a patterned string like "[L1L1L1]" for one target or generating extremely long, patterned numeric strings for others. Lengthy, seemingly gibberish character strings were also common from models like `LLaVA-7b`. This unpredictability underscores a lack of robust and stable visual feature extraction and interpretation, contrasting with human visual processing, which tends towards predictable errors based on similarity.