HueManity: Probing Fine-Grained Visual Perception in MLLMs

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

Abstract

Multimodal Large Language Models (MLLMs) demonstrate strong high-level visual reasoning, yet their foundational understanding of nuanced perceptual details is often overlooked by existing evaluations. To address this, we introduce HueManity, a novel benchmark specifically designed to assess this crucial dimension of MLLM visual understanding. HueManity comprises 83,850 Ishihara-style images with embedded alphanumeric strings, challenging models on precise pattern recognition - a fundamental aspect of visual understanding. Our evaluation of nine MLLMs reveals a profound performance deficit: the best-performing model achieved only 33.6% accuracy on an 'easy' numeric task and 3% on a 'hard' alphanumeric task. This starkly contrasts with human (100% numeric, 95.6% alphanumeric) and fine-tuned ResNet50 (96.5% numeric, 94.5% alphanumeric) performance. These findings uncover a critical gap in MLLMs' fine-grained visual understanding, a limitation not apparent through conventional high-level assessments. HueManity offers a new paradigm for evaluating this specific type of model understanding. We open-source the dataset and code to foster research towards robust perception in MLLMs.

Code: https://github.com/rynaa/huemanity
Dataset: https://huggingface.co/datasets/Jayant-Sravan/HueManity

1. Introduction

Recent advances in 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; 2023; Anthropic, 2025) have enabled sophisticated integration of visual and textual information, leading to strong performance in tasks like image description (Dong et al., 2024; Fu et al., 2024a) and visual question answering (Weng et al., 2025; Chen et al., 2025; Kuang et al., 2024). This success is largely attributed to pre-training on vast image-text datasets, fostering high-level semantic understanding (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).

However, existing evaluations of MLLMs primarily focus on these conceptual capabilities, neglecting their fine-grained perceptual acuity (Bai et al., 2023b; Li et al., 2024; 2023a; Xu et al., 2024; Yin et al., 2023; Liu et al., 2023). This paper addresses this gap by introducing HueManity, a benchmark designed to test MLLMs' ability to discern subtle visual details. 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. HueManity uses controlled Ishihara-style stimuli with embedded alphanumeric characters to assess pattern recognition under visual clutter and subtle color/luminance contrasts.

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

To address this gap and facilitate further research in this domain, this paper makes the following specific contributions: (1) **We introduce HueManity**, a new large-scale benchmark of 83,850 Ishihara-inspired alphanumeric images, meticulously designed with 25 carefully curated color pairs for systematic challenge. (2) We present a **comprehensive evaluation of nine state-of-the-art MLLMs** on HueManity, which reveals a notable performance gap com-

[*]Equal contribution [1]Google, Mountain View, CA, USA [2]Waymo, Mountain View, CA, USA. Correspondence to: Rynaa Grover <rynaa@google.com>, Jayant Sravan Tamarapalli <jayantsravan@google.com>.

ICML 2025 Workshop on Assessing World Models.

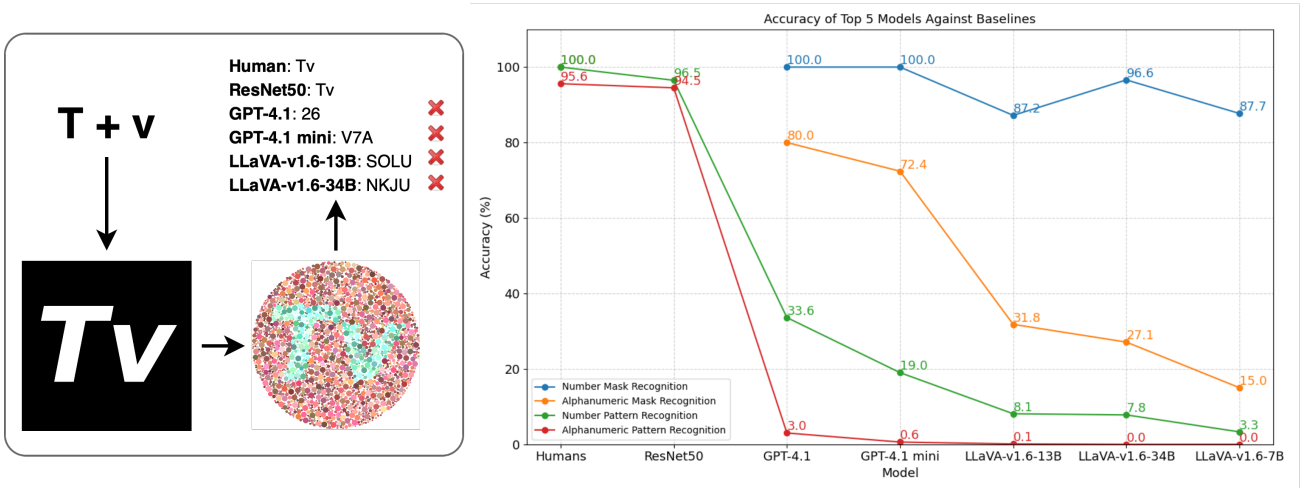

Figure 1. **HueManity** — A new benchmark for MLLM fine-grained visual perception. The pipeline creates a character mask from alphanumeric characters and renders it as an Ishihara-style pattern. While models achieve high accuracy on clear masks, they struggle on the challenging pattern images.

pared to human and fine-tuned ResNet50 baselines, thereby indicating current MLLM limitations in fine-grained perception. (3) We release **open-source code for stimuli generation**, enabling reproducible research and community-driven extensions in this domain.

2. Related Works

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; 2023) have emerged from Large Language Models (LLMs), integrating visual information via modality adaptation layers. Early examples like BLIP-2 (Li et al., 2023b) focused on pre-training for tasks such as Visual Question Answering (VQA), while LLaVA (Liu et al., 2023) advanced instruction tuning with synthetic data. However, MLLMs' success in visual tasks often appears to depend more on strong language capabilities than on fundamental visual perception. HueManity is designed to rigorously evaluate these core, non-linguistic visual skills.

Despite strong global image understanding, MLLMs struggle with fine-grained visual tasks such as precise recognition (Huang & Zhang, 2024; Li et al., 2024). Various benchmarks exist, including TouchStone (Bai et al., 2023b) and LLaVA-Bench (Liu et al., 2023) with their manually annotated visual dialog questions. However, reliance on GPT-based models for evaluation in LLaVA-Bench (Liu et al., 2023), LAMM (Yin et al., 2023), and TouchStone (Bai et al., 2023b) introduces reliability and cost issues. LVLM-eHub (Xu et al., 2024) aggregates benchmarks but remains subjective and expensive due to human annotation. MME and MMBench (Liu et al., 2024b) offer more objective multiple-choice questions for perception and reasoning, with MM-Vet (Yu et al., 2023) extending coverage to OCR and math, but often rely on existing VQA datasets or GPT-generated questions. SEED-Bench (Li et al., 2024; 2023a) provides 24,000 human-annotated questions but remains simple. Blink (Fu et al., 2024b) attempts holistic evaluation across 14 perception tasks but does not assess task combinations and shows high model accuracies, indicating less challenge.

HueManity provides a scalable, objective, and reliable methodology through procedural generation and exact-match evaluation, distinct from more subjective or resource-intensive techniques.

3. Data Creation

The HueManity dataset contains 83,850 images, each with a two-character alphanumeric string with its ground truth label and full generation parameters. We excluded visually ambiguous characters ('l', 'I', 'J', 'O') and combinations starting with '0' to prevent prediction conflicts. The full dataset was generated using 25 carefully chosen color pairs.

For our MLLM evaluations, we sampled two distinct subsets, each with 1,000 randomly selected images, due to time and API cost constraints:

- **Number Recognition Set (Easier Task)**: This subset contains images with only two-digit numeric strings (e.g., 17, 83, 05).
- **Text Recognition Set (Harder Task)**: This subset includes diverse two-character alphanumeric strings (e.g., A7, b9, XG).

3.1. Text Mask Generator

First, we create a 900x900 pixel binary text mask for each two-character string. These masks, rendered in white on a black background using Pygame, utilize the DejaVu Sans font (size 550, bold and italic) to ensure readability and adequate character width for dot-based rendering (Figure 1).

3.2. Ishihara-Style Pattern Generation

Our pattern generator, adapted from an open-source Pygame project[1], iteratively populates the image with non-overlapping circles. Over 30,000 iterations, the generator randomly places circles, computes their maximum non-colliding radius (4-15 pixels), and assigns an initial color (foreground or background) based on whether the circle's center falls within the character mask. This initial color then undergoes three randomized transformations: a gradient shift towards the other pair color, an RGB color shift (range [-30, +30]), and an RGB lightness scaling (factor 0.66 - 1.5). These transformed circles are then rendered, resulting in the final dense Ishihara-style pattern (Figure 1).

3.3. Color Pairs Selection

We meticulously selected 25 distinct foreground-background color pairs for the HueManity stimuli through a multi-stage process. This procedure involved both quantitative CIEDE2000 (Luo et al., 2001) analysis and extensive manual verification, balancing perceptual challenge with human legibility (Appendix D).

4. Experiments

We evaluated nine state-of-the-art Multimodal Large Language Models (MLLMs), including both commercial APIs and open-source models.

Our evaluation focused on two tasks: numerical recognition (digits only) and alphanumeric recognition (digits and letters). For each task, we used a subset of 1,000 two-character strings and their generated images from HueManity. Models were tested on two types of visual stimuli:

- **HueManity Pattern**: The 1,000 randomly sampled Ishihara-style dot pattern images.
- **Text Masks**: The corresponding 1,000 binary text mask images (white text on a black background, as in Figure 1). This baseline helps distinguish a model's fundamental OCR capabilities from its performance on perceptually challenging dot patterns. The prompts used for both are tasks are mentioned in Appendix A.

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

4.1. Human Performance Evaluations

To establish a human baseline, three adult volunteers with self-reported normal color vision were tested on a representative subset of 100 images for each task (numerical and alphanumeric recognition). Volunteers viewed the 900x900 pixel images in a Google Sheets document and entered responses using the same prompts given to the MLLMs, ensuring direct comparison.

4.2. Traditional Computer Vision Baseline (ResNet50)

As a traditional computer vision baseline, we fine-tuned an ImageNet-pretrained ResNet50 model from the PyTorch vision library. We replaced its standard classification layer with two independent classification heads, treating the task as two separate character recognition problems. The model was fine-tuned for 30 epochs on 2,000 randomly sampled HueManity images (distinct from evaluation sets) using the Adam optimizer ($1e-3$ learning rate) and the sum of cross-entropy losses from both heads. The trained model was evaluated on the same 1,000-image subsets used for MLLM evaluations.

5. Results and Analysis

5.1. Human Performance: A Near-Perfect Baseline

Human evaluators established a crucial baseline, demonstrating exceptionally high accuracy and efficiency on the HueManity benchmark (Table 2). For numerical recognition, volunteers achieved a perfect 100% accuracy, while for the more complex alphanumeric task, they achieved a strong 95.6% average accuracy. Annotator feedback highlighted that minor errors in the alphanumeric task stemmed from differentiating visually similar character forms (e.g., 's', 'c', 'w' upper and lower case variants), not from an inability to perceive characters within the dot patterns. Volunteers also processed images remarkably fast, typically under **one second per image**. These near-perfect and rapid human scores are critical as they:

- **Confirms task solvability** and stimulus clarity for proficient visual systems.
- **Establishes a clear performance ceiling** for machine perception on this perceptually simple task.
- **Highlights MLLM-specific limitations**, indicating struggles in perceptual grouping or recognition rather than task impossibility.
- **Contextualizes machine errors**, showing where MLLM perception diverges starkly from human visual understanding.

5.2. ResNet50 Baseline: Demonstrating Task Learnability

A fine-tuned ResNet50 model provided a strong traditional computer vision baseline, achieving 96.5% on the numerical task and 94.5% on the alphanumeric task (Table 1). This robust performance from a standard convolutional architecture, fine-tuned on a relatively small dataset (2,000 images from HueManity), demonstrates that identifying characters within these dot patterns is fundamentally learnable by established computer vision techniques. Achieving near-human accuracy suggests that the perceptual cues are rich enough for a focused model to learn recognition. This implies the task is not inherently intractable for AI and highlights that MLLM difficulties likely stem from how these larger, more general models process or prioritize fine-grained perceptual information.

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

	Number Task		Alphanumeric Task	
	Mask	**Pattern**	**Mask**	**Pattern**
Humans (average)	100%	100%	100%	95.6%
ResNet50	-	96.5%	-	94.5%
API-based models				
GPT-4.1 mini	100%	19.0%	72.4%	0.6%
GPT-4.1	100%	33.6%	80%	3.0%
Claude 3.7 Sonnet	100%	0.4%	82.2%	0%
Open-source models				
LLaVA-v1.6-7B	87.7%	3.3%	15%	0%
LLaVA-v1.6-13B	87.2%	8.1%	31.8%	0.1%
LLaVA-v1.6-34B	96.6%	7.8%	27.1%	0%
Mistral-small3.1-24b	100%	0.1%	58.7%	0%
Qwen VL Max	100%	0.2%	83.5%	0%
Pixtral	100%	1%	65.8%	1.8%

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

The performance of the nine evaluated MLLMs on HueManity sharply contrasts with the near-perfect accuracies of both humans and the ResNet50 baseline (Table 1). All MLLMs consistently struggled, with the best achieving only 33.6% on the numeric task and a mere 3% on the alphanumeric task. This shocking underperformance on demonstrably solvable tasks signals a critical gap in current MLLMs' fine-grained visual perception.

Several characteristics inherent to the current design and training paradigms of many MLLMs may contribute to their observed difficulties on tasks demanding nuanced visual perception.

- **Semantic Optimization, Pipeline Bottleneck**: MLLM vision encoders (*e.g.*, ViT variants (Liu et al., 2023; 2024a; Agrawal et al., 2024; Bai et al., 2023a; 2025), often optimized for semantic information and global scene context, might unintentionally de-emphasize or lose fine-grained local details crucial for these tasks (e.g., subtle color shifts defining patterns in cluttered backgrounds). The projection layers connecting vision encoders to language models (Liu et al., 2023; 2024a; Agrawal et al., 2024; Bai et al., 2023a; 2025) could also act as an information bottleneck, abstracting or losing precise, high-resolution feature distinctions.

- **Impact of Pre-Training Paradigms on Foundational Visual Acuity**: MLLMs are primarily pre-trained on vast web-scale corpora with image-text pairings (Liu et al., 2023; 2024a; Agrawal et al., 2024; Bai et al., 2023a; 2025; Achiam et al., 2023). While fostering semantic alignment and contextual understanding, these datasets may not adequately represent stimuli requiring intensive perceptual organization based purely on low-level visual features without strong linguistic or object-based anchors. Consequently, MLLMs might lack specialized visual routines needed for tasks like grouping elements based on subtle shared properties (*e.g.*, color similarity). Their success often relies on powerful integrated LLMs for conceptual interpretation and reasoning (Achiam et al., 2023; Anthropic, 2025; Liu et al., 2023; Gong et al., 2023; Jiang et al., 2023). This reliance is less effective when the core challenge demands direct, bottom-up visual processing and pattern extraction rather than semantic inference. Such tasks require a foundational visual acuity that may not be a primary emergent outcome of training focused on multimodal semantics and instruction following.

6. Conclusion and Future Directions

Our HueManity benchmark (83,850 Ishihara-style images) assesses MLLM fine-grained visual perception, uncovering critical limitations. Evaluations revealed a stark MLLM performance gap: top models scored as low as 3% on challenging alphanumeric tasks (and only 33.6% on easier numeric ones), far underperforming human and ResNet50 baselines in discerning patterns amidst visual noise. HueManity will be open-sourced to spur further research. Future work should target novel MLLM architectures, data, and training objectives to improve foundational visual acuity.

7. Limitations

HueManity focuses on a specific task of identifying two-character alphanumeric strings, and its direct generalization to the full spectrum of real-world fine-grained challenges requires further investigation. The dataset primarily explores color and basic character forms; future work could expand to texture, orientation, or motion.

References

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

Agrawal, P., Antoniak, S., Hanna, E. B., Bout, B., Chaplot, D., Chudnovsky, J., Costa, D., Monicault, B. D., Garg, S., Gervet, T., Ghosh, S., Héliou, A., Jacob, P., Jiang, A. Q., Khandelwal, K., Lacroix, T., Lample, G., Casas, D. L., Lavril, T., Scao, T. L., Lo, A., Marshall, W., Martin, L., Mensch, A., Muddireddy, P., Nemychnikova, V., Pellat, M., Platen, P. V., Raghuraman, N., Rozière, B., Sablayrolles, A., Saulnier, L., Sauvestre, R., Shang, W., Soletskyi, R., Stewart, L., Stock, P., Studnia, J., Subramanian, S., Vaze, S., Wang, T., and Yang, S. Pixtral 12b, 2024. URL https://arxiv.org/abs/2410.07073.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., and Simonyan, K. Flamingo: a visual language model for few-shot learning, 2022. URL https://arxiv.org/abs/2204.14198.

Anthropic. Claude 3.5 sonnet, 2025. URL https://claude.ai/. Accessed May 18, 2025.

Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., and Zhou, J. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023a.

Bai, S., Yang, S., Bai, J., Wang, P., Zhang, X., Lin, J., Wang, X., Zhou, C., and Zhou, J. Touchstone: Evaluating vision-language models by language models. *arXiv preprint arXiv:2308.16890*, 2023b.

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., and Lin, J. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Chen, B., Khare, A., Kumar, G., Akula, A., and Narayana, P. Seeing beyond: Enhancing visual question answering with multi-modal retrieval. In Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B. D., Schockaert, S., Darwish, K., and Agarwal, A. (eds.), *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pp. 410–421, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-industry.35/.

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

Dong, R., Han, C., Peng, Y., Qi, Z., Ge, Z., Yang, J., Zhao, L., Sun, J., Zhou, H., Wei, H., Kong, X., Zhang, X., Ma, K., and Yi, L. Dreamllm: Synergistic multimodal comprehension and creation, 2024. URL https://arxiv.org/abs/2309.11499.

Fu, T.-J., Hu, W., Du, X., Wang, W. Y., Yang, Y., and Gan, Z. Guiding instruction-based image editing via multimodal large language models, 2024a. URL https://arxiv.org/abs/2309.17102.

Fu, X., Hu, Y., Li, B., Feng, Y., Wang, H., Lin, X., Roth, D., Smith, N. A., Ma, W.-C., and Krishna, R. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2024b.

Gong, T., Lyu, C., Zhang, S., Wang, Y., Zheng, M., Zhao, Q., Liu, K., Zhang, W., Luo, P., and Chen, K. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*, 2023.

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

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q. V., Sung, Y., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision, 2021. URL https://arxiv.org/abs/2102.05918.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Kuang, J., Xie, J., Luo, H., Li, R., Xu, Z., Cheng, X., Li, Y., Lin, X., and Shen, Y. Natural language understanding and inference with mllm in visual question answering: A survey, 2024. URL https://arxiv.org/abs/2411.17558.

Li, B., Wang, R., Wang, G., Ge, Y., Ge, Y., and Shan, Y. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023a.

Li, B., Ge, Y., Ge, Y., Wang, G., Wang, R., Zhang, R., and Shan, Y. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13299–13308, 2024.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023b.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024a.

Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024b.

Luo, M., Cui, G., and Rigg, B. The development of the cie 2000 colour-difference formula: Ciede2000. *Color Research Application*, 26:340 – 350, 10 2001. doi: 10.1002/col.1049.

Pham, H., Dai, Z., Ghiasi, G., Kawaguchi, K., Liu, H., Yu, A. W., Yu, J., Chen, Y.-T., Luong, M.-T., Wu, Y., Tan, M., and Le, Q. V. Combined scaling for zero-shot transfer learning, 2023. URL https://arxiv.org/abs/2111.10050.

Qi, D., Su, L., Song, J., Cui, E., Bharti, T., and Sacheti, A. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data, 2020. URL https://arxiv.org/abs/2001.07966.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., and Jitsev, J. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. URL https://arxiv.org/abs/2210.08402.

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

Weng, W., Zhu, J., Meng, X., Zhang, H., Zhang, R., and Yuan, C. Learning to compress contexts for efficient knowledge-based visual question answering, 2025. URL https://arxiv.org/abs/2409.07331.

Xu, P., Shao, W., Zhang, K., Gao, P., Liu, S., Lei, M., Meng, F., Huang, S., Qiao, Y., and Luo, P. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Yin, Z., Wang, J., Cao, J., Shi, Z., Liu, D., Li, M., Huang, X., Wang, Z., Sheng, L., Bai, L., et al. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *Advances in Neural Information Processing Systems*, 36:26650–26685, 2023.

Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., and Wang, L. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., and Beyer, L. Lit: Zero-shot transfer with locked-image text tuning, 2022. URL https://arxiv.org/abs/2111.07991.

Table 2. Human evaluation accuracies on 100-image subsets for both number and alphanumeric pattern recognition tasks.

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

A. MLLM Prompts

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]"

B. 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.

B.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.

B.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.

B.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.

C. A Brief Discussion on CIEDE2000 Color Difference

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

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

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 2. 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.

D. 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

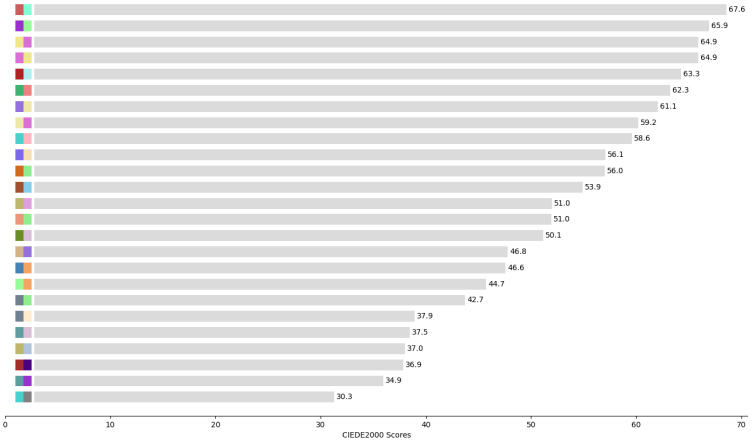

Figure 2. Distribution of CIEDE2000 color difference scores for the 25 selected foreground-background color pairs utilized in the HueManity benchmark.

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. 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.

E. Usage of Generative AI tools

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