PROBING PERCEPTUAL CONSTANCY IN LARGE VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Perceptual constancy is the ability to maintain stable perceptions of objects despite changes in sensory input, such as variations in distance, angle, or lighting. This ability is crucial for visual understanding in a dynamic world. Here, we explored such ability in current Vision Language Models (VLMs). In this study, we evaluated 155 VLMs using 236 experiments across three domains: color, size, and shape constancy. The experiments included single-image and video adaptations of classic cognitive tasks, along with novel tasks in in-the-wild conditions. We found significant variability in VLM performance across these domains, with model performance in shape constancy clearly dissociated from that of color and size constancy.

1 INTRODUCTION

Perceptual constancy is the ability to perceive the properties of environmental objects as stable despite variations in external conditions. While this phenomenon exists across multiple sensory modalities, including auditory and tactile perception, it is particularly well-studied in vision. The human visual system demonstrates remarkable constancy, maintaining stable object perception despite changes in distance, viewing angle, lighting (Epstein, 1977; Walsh and Kulikowski, 1998). This stability allows humans to perceive objects quickly and accurately, even in complex and dynamic scenes, ensuring seamless interaction with the surrounding world.

The study of perceptual constancy dates back to the 19th century, with early theories proposing that it emerges from the brain's inference process—where stable perception is achieved by integrating empirical knowledge and environmental cues (Von Helmholtz, 1867). Neuroscientific research has since demonstrated that visual information is processed hierarchically, from the retina to the visual cortex (e.g., V1, V4, IT cortex), to maintain perceptual stability across changing conditions (DiCarlo et al., 2012). For example, neurons in the IT cortex consistently encode color regardless of illumination, suggesting that different levels of the visual system progressively compensate for environmental variations to preserve object properties (Tanaka, 1996). This intricate mechanism enables humans to navigate dynamic environments, recognize objects across varying perspectives, and make accurate judgments in complex real-world scenarios. Similarly, instilling AI models with human-like perceptual constancy is crucial for developing robust vision-language models (VLMs) that can perform reliably under diverse and unpredictable conditions. Given the rapid progress of VLMs, perceptual constancy therefore appears to be a key benchmark for assessing their cognitive capabilities and identifying their limitations.

Perceptual constancy is essential for overcoming visual ambiguities in everyday life. Color constancy allows us to recognize objects consistently despite changes in lighting—a critical ability for autonomous systems in varying illumination conditions, such as self-driving cars interpreting traffic signals at different times of day (Land and McCann, 1971). Size constancy ensures that objects appear the same size regardless of distance, facilitating spatial awareness and depth perception. This is particularly important for AI applications in robotics, where accurate size perception is necessary for grasping and object manipulation (Carlson, 2010). Shape constancy enables stable object recognition across different viewpoints, supporting real-world tasks such as facial recognition, medical imaging, and augmented reality applications (Sternberg, 2006). Without these constancy mechanisms, AI systems may struggle with perceptual inconsistencies, leading to unreliable in-the-wild performance that lacks adaptability and precision.

Figure 1: **Sample Tasks from the Three Evaluation Dimensions of ConstancyBench.** Example model performance from GPT-4o is presented.

To systematically assess the perceptual constancy capabilities of VLMs, we leveraged the ConstancyBench from the **CoreCognition** benchmark (Li et al., 2024b), formulating a dataset comprising 253 cognitive experiments. These experiments specifically examine the three primary dimensions of perceptual constancy: color, size, and shape. By testing 155 VLMs on these fundamental perceptual tasks, we aim to uncover the extent to which current models exhibit perceptual stability and where they fall short. Understanding these limitations is essential for advancing AI's ability to process visual information in a way that aligns more closely with human perception, ultimately enabling more reliable and adaptable AI systems in real-world applications.

2 METHODS

2.1 EXPERIMENT DESIGN

Perceptual constancy refers to the human ability to perceive object properties consistently across varying environmental conditions, here focusing on three key domains: color, size, and shape constancy. These three aspects were selected to capture the fundamental principles of constancy in visual perception, enabling stable object recognition despite changes in lighting, distance, and viewing angle. Below, we provided explanations for each domain.

2.1.1 COLOR CONSTANCY

Color constancy is an important feature of the human visual system. It allows us to perceive the color of objects consistently under different light conditions. A common example is a white wall appearing in different shades under different lighting conditions. However, the human visual system still perceives it as white rather than the wall's actual color has changed. This occurs because the visual system can separate an object's true color from the influence of lighting condition, thereby maintaining stable color perception (Jameson and Hurvich, 1989). Evaluating color constancy can reveal whether VLMs can truly understand an object's intrinsic color rather than merely relying on color patterns in the training data.

2.1.2 Size Constancy

Size constancy refers to the perception of an object's size as stable, even when its retinal image changes due to variations in distance (Sperandio and Chouinard, 2015). For example, a distant car appears just as large as a nearby one, despite the difference in retinal projection. This stability is crucial for spatial awareness, depth perception, and navigation. Assessing this phenomenon in VLMs can determine whether they truly grasp the spatial properties of objects in the dynamic environment.

2.1.3 Shape Constancy

Shape constancy allows us to recognize objects as having the same shape, even when viewed from different angles (Rock, 1973). A round plate, for example, may project an elliptical image when seen obliquely, yet we still perceive it as circular. This perceptual stability relies on depth cues, prior experience, and contextual information. Shape constancy is fundamental to object recognition and spatial reasoning, allowing for accurate identification across perspectives. Research suggests humans achieve this by comparing novel views to stored shape representations (Tarr, 1995). Evaluating shape constancy in VLMs can reveal whether they truly understand an object's three-dimensional form or rely on fixed representations contingent to certain viewpoints.

2.2 Examined Vision Language Models

Recent advances in multi-modal learning have been driven by the unified modeling of visual and textual modalities using transformers (Li et al., 2019; Xu et al., 2023; Tan and Bansal, 2019; Alayrac et al., 2022; Radford et al., 2021). With the rise of large language models (LLMs), state-of-the-art (SOTA) multi-modal LLMs (MLLMs) (Liu et al., 2024; Li et al., 2023b) adopt open-source LLMs (Touvron et al., 2023; Peng et al., 2023; Jiang et al., 2023) and align visual features to the LLM embedding space (Li et al., 2023a; Fu et al., 2023; Wu and Xie, 2024; Xu et al., 2024; Shao et al., 2024; Li et al., 2022; 2025; Brown et al., 2020; Achiam et al., 2023; Bai et al., 2023; Jaech et al., 2024; Zhang et al., 2025; 2024). Progressively, MLLMs have demonstrated competitive performance in complex tasks involving high-level perception and reasoning (Li et al., 2024a; Gemini, 2023; Fu et al., 2023; OpenAI, 2023), such as spatial reasoning (Chen et al., 2024; Cai et al., 2024), character recognition (Mori et al., 1999), scene understanding (Cordts et al., 2016; Wang et al., 2023; Li et al., 2023c; Chen et al., 2017), action recognition (Jhuang et al., 2013; Herath et al., 2017) and prediction (Lan et al., 2014; Kong and Fu, 2022), reaching near-human performance.

We evaluated a total of 155 models for perceptual constancy analysis. Model performances are reported across three domains—color, size, and shape constancy. Each model was tested under a zero-shot setting, generating answers and textual explanations for each experimental prompt. Model size data and architecture type were recorded for correlation analysis.

2.3 Data Sources

The dataset includes six types of data sources: photographs (videos) taken, images (videos) from past classic cognitive psychology experiments, movies (animations), hand-drawn works, and AI-generated images (videos). The distribution is as follows:

1. There are 223 instances of photographs (videos);
2. There are 25 instances of images (videos) from classic cognitive psychology experiments;
3. There are 5 AI-generated images, involving standard 3D geometric shapes only.

Some photographs and video materials have undergone post-processing for experimental purposes.

3 Results

3.1 General Results

Across 155 evaluated Vision-Language Models (VLMs), perceptual constancy performance varies substantially both across and within architectures. Accuracy spans a wide range—from below 0.20

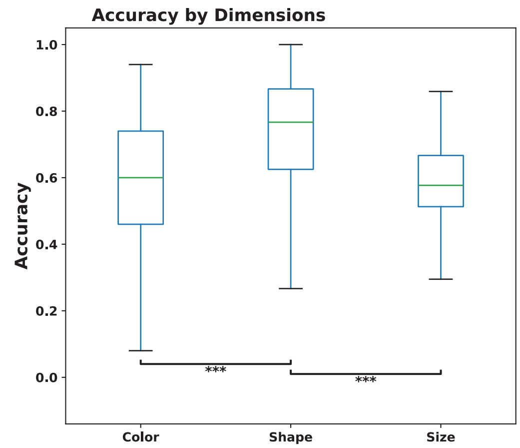


Figure 2: Bar plots show accuracy scores across vision–language models for color ($n = 153$), shape ($n = 152$), and size ($n = 149$) constancy tasks after outlier removal (Color: 0.588 ± 0.185; Shape: 0.723 ± 0.170; Size: 0.584 ± 0.123). Horizontal bars with asterisks indicate statistical significance from post-hoc Tukey HSD tests ($*\, p < 0.001$; ns = not significant, $p > 0.05$).

in lightweight open-source models to above 0.90 in frontier systems such as GPT-4o and Gemini 1.5 Pro—highlighting a striking heterogeneity in perceptual stability. This variability underscores that perceptual constancy remains an emergent rather than universal capability: while certain large models exhibit localized robustness under specific transformations, no current architecture demonstrates consistent generalization across all constancy dimensions.

3.2 Domain-wise Differences and Statistical Analysis

Figure 2 summarizes model performance across the three perceptual constancy domains. VLMs demonstrate a clear performance hierarchy, achieving highest accuracy on shape constancy ($M = 0.723$, $SD = 0.170$), followed by color constancy ($M = 0.588$, $SD = 0.185$) and size constancy ($M = 0.584$, $SD = 0.123$). Statistical analysis confirmed these domain-specific differences through a one-way ANOVA revealing a robust main effect of perceptual domain on accuracy ($F(2, 451) = 36.49$, $p = 2.05 \times 10^{-15}$, $\eta^2 = 0.139$). Post-hoc Tukey HSD comparisons demonstrated that shape constancy performance significantly exceeded both color ($p < 0.001$, Cohen's $d = 0.78$) and size constancy ($p < 0.001$, Cohen's $d = 0.89$), while color and size constancy showed statistically equivalent performance ($p = 0.976$, Cohen's $d = 0.02$).

These findings suggest shape constancy exploits simple geometric shortcuts, making it the most tractable domain for current VLMs. In contrast, color constancy requires high-dimensional photometric representations, while size constancy demands 3D world representation—explaining why shape dominates while color and size remain computationally challenging across models.

3.3 Relationship Between Model Performance and Model Size

To evaluate how architectural scale influences perceptual constancy, we examined the relationship between model performance and parameter count (Figure 3). Across 151 VLMs for which parameter estimates were available, regression analyses revealed a robust and consistent scaling effect. Overall accuracy increased significantly with model size ($R^2 = 0.2804$, $t = 7.62$, $p = 2.74 \times 10^{-12}$), following the log-linear relation $y = 0.1200x + 0.4766$ where $x = \log_{10}(\text{param})$. Similar positive correlations were observed in all three perceptual constancy dimensions: color constancy ($R^2 = 0.0973$, $t = 4.01$, $p = 9.66 \times 10^{-5}$, $y = 0.0993x + 0.4888$), shape constancy ($R^2 = 0.1148$, $t = 4.40$, $p = 2.08 \times 10^{-5}$, $y = 0.1017x + 0.6171$), and size constancy ($R^2 = 0.3080$, $t = 8.14$, $p = 1.42 \times 10^{-13}$, $y = 0.1301x + 0.4457$).

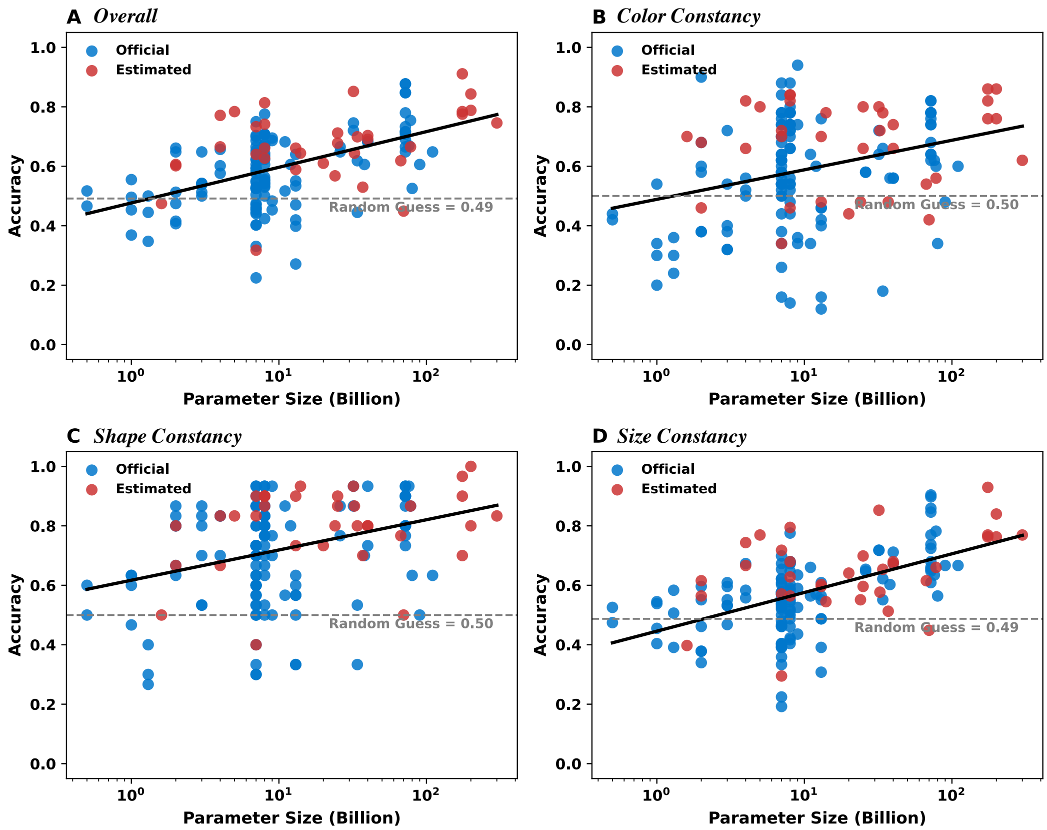


Figure 3: **Relationship between model size and performance.** Larger models tend to perform better in perceptual constancy tasks.

These results collectively demonstrate that larger VLMs tend to achieve systematically higher perceptual stability across all constancy domains. The scaling relationships are not only statistically significant but also qualitatively consistent: as model capacity increases, representations become progressively more invariant to color, size, and shape perturbations. Among the three dimensions, the strongest scaling effect was observed in **size constancy**, indicating that geometric and spatial reasoning benefits most directly from expanded model capacity. In contrast, **color constancy** showed a weaker but still reliable dependence on size, suggesting that photometric invariance emerges more gradually with scale. Taken together, these findings support a unified scaling law of perceptual stability: *the larger the model, the greater its resilience to perceptual variation*—a trend that spans all examined dimensions without evidence of domain-specific exceptions.

3.4 Item Response Theory (IRT) Analysis

To examine the latent structure underlying VLM performance across perceptual constancy tasks, we fitted a two-parameter logistic (2PL) Item Response Theory (IRT) model to all evaluated items. This framework characterizes each item by its discrimination parameter (a), which quantifies how sensitively the item differentiates between higher- and lower-ability models, and its difficulty parameter (b), which represents the ability level required to achieve a 50% success probability.

Across the three perceptual domains, clear structural distinctions emerged (Figure 4A and Figure 4B). **Color** items exhibited the highest mean discrimination ($\bar{a} = 2.25$, $SD = 1.62$) and moderate difficulty ($\bar{b} = -0.58$, $SD = 1.32$), indicating that color-related tasks sharply separate stronger from weaker models, but require only moderate ability for success. **Shape** items showed lower discrimination ($\bar{a} = 1.33$, $SD = 0.84$) combined with the lowest mean difficulty ($\bar{b} = -1.27$, $SD = 1.08$), reflecting tasks that most models could solve reliably and con-

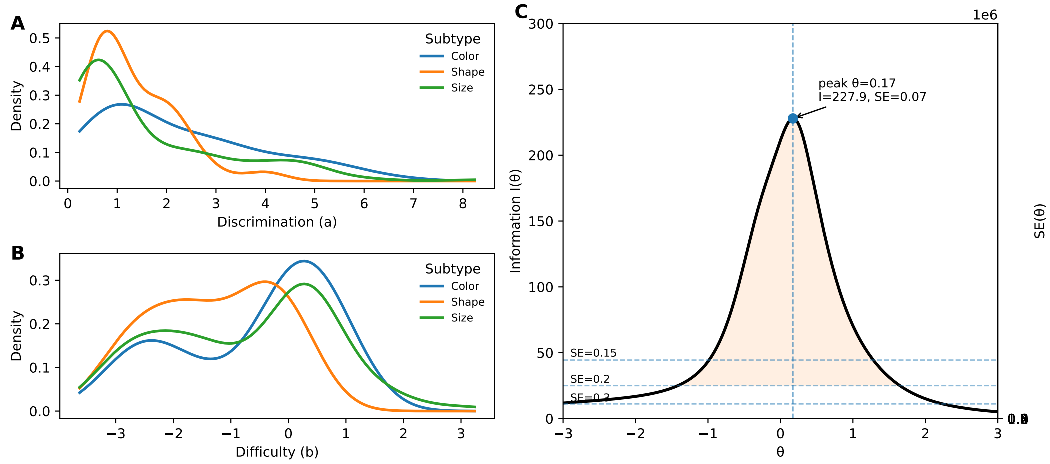


Figure 4: **Item Response Theory (IRT) analysis of model-level performance.** Each point represents a single item parameterized by discrimination (a) and difficulty (b). Shape-constancy items cluster in regions of lower difficulty and moderate discrimination, whereas color and size items show wider dispersion, reflecting greater variability in perceptual demands and model sensitivity.

sistently—consistent with their high observed pass rate (0.76). In contrast, **Size** items demonstrated intermediate discrimination ($\bar{a} = 1.59$, $SD = 1.56$) and the broadest difficulty distribution ($\bar{b} = -0.72$, $SD = 1.47$), suggesting greater heterogeneity in spatial reasoning demands and item complexity within this domain.

The global test information curve (Figure 4C) peaks around $\theta \approx 0$, indicating maximal measurement precision for models of average ability, with standard error (SE) below 0.20 across a wide ability range ($-1.2 < \theta < +1.1$). Together, these results reveal a well-calibrated hierarchy of perceptual challenges: shape constancy tasks anchor the lower end of the difficulty continuum and provide stable measurement of baseline competence, while color and size constancy tasks contribute discriminative power at moderate to high ability levels. This latent structure mirrors behavioral accuracy trends.

4 Discussion

This study introduces **ConstancyBench**, a large-scale benchmark for evaluating perceptual constancy in 155 Vision-Language Models (VLMs) across three fundamental domains—color, size, and shape constancy. Our results reveal that perceptual constancy is not a uniform emergent ability but a stratified one.

Hierarchical emergence of constancy. The clear dissociation between shape versus color and size constancy supports the existence of a representational hierarchy in how VLMs internalize perceptual stability. Shape constancy—rooted in geometric invariance and low-level topology—appears to emerge from the structural regularities encoded within early convolutional or transformer layers. Such mechanisms correspond to what has been termed *minimal constancy*: the capacity to preserve stable representations through accumulated perceptual regularities rather than explicit reasoning about context (Bradley, 2008; Buccella, 2021; Buccella and Chemero, 2022). By contrast, color and size constancy require broader contextual integration, including illumination estimation, depth reasoning, and perspective correction—abilities that depend more heavily on multimodal fusion and scene-level abstraction. This hierarchical division parallels the cognitive progression observed in human perception, where geometric stability precedes the development of higher-order contextual invariances.

Scaling dynamics and emergent thresholds. Our scaling analyses reveal that improvements in constancy performance are systematic yet nonlinear. Regression trends indicate that all three domains benefit from increased model capacity, but the relationship is strongest for size constancy, which shows a steep performance gain beyond approximately 10–13 billion parameters. This pattern echoes the "emergent threshold" phenomena widely observed in large language models, where complex behaviors—such as compositional reasoning or in-context adaptation—abruptly appear once sufficient representational depth is achieved. Analogously, perceptual constancy in VLMs may follow a *phase-transition dynamic*, where representational stability reorganizes once the network's capacity surpasses a critical complexity threshold. This finding bridges cognitive theories of perceptual maturation with the scaling laws of modern foundation models.

Cognitive interpretation and representational implications. From a cognitive perspective, the observed hierarchy mirrors developmental patterns in human perception: infants acquire shape constancy early through sensorimotor interaction, whereas color and size constancy emerge later as the visual system integrates cross-modal cues. VLMs seem to recapitulate this trajectory at the algorithmic level—progressing from local geometric consistency toward contextual and relational invariance. However, their mechanism of acquisition is fundamentally different: rather than inferential generalization from embodied experience, these models achieve constancy via statistical regularities embedded in massive multimodal datasets. This raises an important conceptual question: do VLMs exhibit *functional constancy*—genuine invariance to physical transformations—or merely *correlational constancy*, a byproduct of data coverage and alignment? Addressing this distinction will be key to evaluating whether these systems possess perceptual understanding or only statistical resemblance to it.

Toward mechanistic understanding and future directions. Mechanistically, probing how constancy-related invariances are encoded across visual and cross-attentional layers could reveal whether these patterns reflect bottom-up feature persistence or top-down integration. Architecturally, scaling alone may be insufficient: inductive biases such as explicit 3D priors, lighting normalization, or causal simulation modules may be required to support genuine invariance. Extending **ConstancyBench** to video, embodied, and generative contexts could further assess how temporal continuity and physical interaction contribute to perceptual stability. Such extensions would enable evaluation of *causal perceptual reasoning*—the ability not only to maintain perceptual constancy but also to predict and explain variation under changing conditions. Ultimately, perceptual constancy offers a powerful diagnostic lens: it operationalizes the transition from mere perception to world modeling, providing a measurable bridge between cognitive theory and the representational dynamics of modern AI.

5 CONCLUSION

This work establishes **ConstancyBench** as the first systematic framework for probing perceptual constancy in large Vision-Language Models (VLMs). Through 236 controlled experiments spanning color, size, and shape constancy, we reveal that current VLMs exhibit a stratified rather than unified form of perceptual stability. Shape constancy emerges robustly even in smaller models, whereas color and size constancy depend strongly on scale and multimodal integration capacity. Our findings uncover a hierarchical pattern of perceptual emergence—one that mirrors developmental trajectories in human perception—and demonstrate that perceptual robustness follows a quantifiable scaling law across modern multimodal systems.

Beyond benchmarking, this study suggests perceptual constancy as a cognitive lens for evaluating the *world-modeling fidelity* of foundation models. Future work should examine how such invariances arise within model representations, whether through implicit statistical regularities or explicit geometric grounding, and extend these analyses to embodied and dynamic settings. Understanding and mechanistically modeling perceptual constancy will be central to bridging the gap between visual perception and true physical understanding in artificial intelligence.

REFERENCES

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.

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.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Peter Bradley. Constancy, categories and bayes: A new approach to representational theories of color constancy. *Philosophical Psychology*, 21(5):601–627, 2008.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Alessandra Buccella. The problem of perceptual invariance. *Synthese*, 199(5):13883–13905, 2021.

Alessandra Buccella and Anthony Chemero. Reconsidering perceptual constancy. *Philosophical Psychology*, 35(7):1057–1071, 2022.

Wenxiao Cai, Yaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. *arXiv preprint arXiv:2406.13642*, 2024.

Neil Carlson. *Psychology: The Science of Behavior*. Pearson Canada Inc., Toronto, ON, Canada, 4th canadian edition, 2010. ISBN 978-0-205-64524-4.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465, 2024.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4): 834–848, 2017.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

James J. DiCarlo, Davide Zoccolan, and Nicole C. Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012. doi: 10.1016/j.neuron.2012.01.010.

William Epstein. *Stability and Constancy in Visual Perception: Mechanisms and Processes*. Wiley-Interscience, New York, 1977.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv: 2306.13394*, 2023.

Gemini. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv: 2312.11805*, 2023.

Samitha Herath, Mehrtash Harandi, and Fatih Porikli. Going deeper into action recognition: A survey. *Image and vision computing*, 60:4–21, 2017.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

D. Jameson and L. M. Hurvich. Essay concerning color constancy. *Annual Review of Psychology*, 40:1–22, 1989. doi: 10.1146/annurev.psych.40.1.1.

Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. Towards understanding action recognition. *In Proceedings of the IEEE international conference on computer vision*, pages 3192–3199, 2013.

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. *arXiv preprint arXiv: 2310.06825*, 2023.

Yu Kong and Yun Fu. Human action recognition and prediction: A survey. *International Journal of Computer Vision*, 130(5):1366–1401, 2022.

Tian Lan, Tsung-Chuan Chen, and Silvio Savarese. A hierarchical representation for future action prediction. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part III 13*, pages 689–704. Springer, 2014.

Edwin Land and John McCann. Lightness and retinex theory. *Journal of the Optical Society of America*, 61:1–11, 01 1971. doi: 10.1364/JOSA.61.000001.

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, 2024a.

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, 2023a.

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 *ICML*, 2023b.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

Yijiang Li, Wentian Cai, Ying Gao, Chengming Li, and Xiping Hu. More than encoder: Introducing transformer decoder to upsample. In *2022 IEEE international conference on bioinformatics and biomedicine (BIBM)*, pages 1597–1602. IEEE, 2022.

Yijiang Li, Xinjiang Wang, Lihe Yang, Litong Feng, Wayne Zhang, and Ying Gao. Diverse cotraining makes strong semi-supervised segmentor. *arXiv preprint arXiv:2308.09281*, 2023c.

Yijiang Li, Qingying Gao, Tianwei Zhao, Bingyang Wang, Haoran Sun, Haiyun Lyu, Robert D Hawkins, Nuno Vasconcelos, Tal Golan, Dezhi Luo, et al. Core knowledge deficits in multimodal language models. *arXiv preprint arXiv:2410.10855*, 2024b.

Yijiang Li, Genpei Zhang, Jiacheng Cheng, Yi Li, Xiaojun Shan, Dashan Gao, Jiancheng Lyu, Yuan Li, Ning Bi, and Nuno Vasconcelos. Egoprivacy: What your first-person camera says about you? *arXiv preprint arXiv:2506.12258*, 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

Shunji Mori, Hirobumi Nishida, and Hiromitsu Yamada. *Optical character recognition*. John Wiley & Sons, Inc., 1999.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv: 2303.08774*, 2023.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023.

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. *arXiv preprint arXiv: 2103.00020*, 2021.

Irvin Rock. *Orientation and Form*. Academic Press, New York, 1973. ISBN 978-0125912501. Digitized on 2008-10-30.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models. *arXiv preprint arXiv:2403.16999*, 2024.

Irene Sperandio and Philippe A. Chouinard. The mechanisms of size constancy. *Multisensory Research*, 28(3-4):253–283, 2015. doi: 10.1163/22134808-00002483.

Robert Sternberg. *Cognitive Psychology*. Wadsworth, Cengage Learning, Belmont, CA, 2006. ISBN 978-0-495-50629-4.

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

Keiji Tanaka. Inferotemporal cortex and object vision. *Annual Review of Neuroscience*, 19:109–139, 1996. ISSN 0147-006X. doi: 10.1146/annurev.ne.19.030196.000545. URL https://doi.org/10.1146/annurev.ne.19.030196.000545.

Michael J. Tarr. Rotating objects to recognize them: A case study on the role of viewpoint dependency in the recognition of three-dimensional objects. *Psychonomic Bulletin & Review*, 2:55–82, 1995. doi: 10.3758/BF03214412. URL https://doi.org/10.3758/BF03214412.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Hermann Von Helmholtz. *Handbuch der physiologischen Optik*, volume 9. Voss, 1867.

Vincent Walsh and Janusz Kulikowski. *Perceptual Constancy: Why Things Look as They Do*. Cambridge University Press, Cambridge, 1998.

Xinjiang Wang, Xingyi Yang, Shilong Zhang, Yijiang Li, Litong Feng, Shijie Fang, Chengqi Lyu, Kai Chen, and Wayne Zhang. Consistent-teacher: Towards reducing inconsistent pseudo-targets in semi-supervised object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3240–3249, 2023.

Penghao Wu and Saining Xie. V?: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13084–13094, 2024.

Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024.

Xiao Xu, Chenfei Wu, Shachar Rosenman, Vasudev Lal, Wanxiang Che, and Nan Duan. Bridgetower: Building bridges between encoders in vision-language representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10637–10647, 2023.

Wanpeng Zhang, Zilong Xie, Yicheng Feng, Yijiang Li, Xingrun Xing, Sipeng Zheng, and Zongqing Lu. From pixels to tokens: Byte-pair encoding on quantized visual modalities. *arXiv preprint arXiv:2410.02155*, 2024.

Wanpeng Zhang, Yicheng Feng, Hao Luo, Yijiang Li, Zihao Yue, Sipeng Zheng, and Zongqing Lu. Unified multimodal understanding via byte-pair visual encoding. *arXiv preprint arXiv:2506.23639*, 2025.

A Appendix

A.1 Model accuracy and IRT-derived ability parameters across perceptual dimensions.

Table 1:

Model	Accuracy	θ_{total}	θ_{color}	θ_{shape}	θ_{size}
o1	0.911	1.707	0.820	0.967	0.929
Qwen2.5-VL-72B-Instruct	0.877	1.394	0.780	0.933	0.897
Eagle-X5-13B-Chat	0.640	1.366	0.950	0.900	0.683
Qwen2.5-VL-72B-Instruct_video	0.877	1.361	0.780	0.900	0.904
qwen-vl-max	0.852	1.356	0.800	0.933	0.881
gemini-1.5-pro	0.788	1.307	0.860	0.800	0.763
gemini-1.5-flash	0.814	1.118	0.820	0.900	0.795
Qwen2-VL-72B-Instruct	0.847	1.083	0.760	0.933	0.859
MiniCPM-Llama3-V-2_5	0.742	1.013	0.840	0.900	0.679
Eagle-X5-34B-Chat	0.619	0.983	0.825	0.800	0.698
gpt-4-turbo	0.847	0.979	0.820	0.900	0.846
gpt-4o	0.843	0.902	0.760	1.000	0.840
LLaVA-NeXT-Video-7B_multi_frame	0.403	0.873	0.857	0.562	0.519
Janus-Pro-7B	0.593	0.858	0.975	0.850	0.604
h2ovl-mississippi-1b	0.716	0.844	0.820	0.900	0.647
xgen-mm-phi3-dpo-r-v1.5	0.784	0.816	0.800	0.833	0.769
gemini-1.5-flash-8b	0.775	0.801	0.720	0.867	0.776
Llama-3-LongVILA-8B-512Frames	0.500	0.766	0.780	0.533	0.404
Ovis1.5-Gemma2-9B	0.686	0.756	0.940	0.767	0.590
LLaVA-Video-7B-Qwen2_multi_frame	0.750	0.753	0.880	0.933	0.673
LLaVA-NeXT-Video-7B-DPO_multi_frame	0.445	0.732	0.842	0.600	0.500
Kosmos2	0.475	0.728	0.700	0.500	0.400
Mantis-8B-Fuyu	0.555	0.721	0.780	0.714	0.468
InternVL-Chat-V1-2	0.703	0.720	0.740	0.800	0.673
xgen-mm-phi3-interleave-r-v1.5	0.771	0.712	0.820	0.833	0.744
VideoChat2_stage3_Mistral_7B	0.623	0.690	0.796	0.800	0.538
Llama-3-LongVILA-8B-1024Frames	0.564	0.684	0.880	0.800	0.417
Ovis1.5-Llama3-8B	0.674	0.667	0.780	0.933	0.590
LLaVA-NeXT-Video-32B-Qwen_multi_frame	0.746	0.653	0.720	0.933	0.718
llava_next_mistral_7b	0.665	0.652	0.680	0.867	0.622
LLaVA-Video-7B-Qwen2	0.703	0.610	0.840	0.900	0.622
LLaVA-Video-72B-Qwen2_multi_frame	0.780	0.606	0.740	0.900	0.769
InternVL2_5-78B	0.754	0.581	0.600	0.867	0.782
claude-3-7-sonnet-20250219	0.708	0.571	0.714	0.867	0.757
Llama-3-LongVILA-8B-256Frames	0.542	0.568	0.800	0.733	0.423
InternVL2-8B-MPO-CoT	0.708	0.562	0.740	0.833	0.673
InternVL2-8B-MPO	0.708	0.562	0.740	0.833	0.673
llava-onevision-qwen2-72b-ov-chat-hf	0.708	0.562	0.796	0.733	0.679
llava-onevision-qwen2-72b-ov-hf	0.691	0.561	0.740	0.767	0.673
h2ovl-mississippi-2b	0.513	0.559	0.600	0.633	0.462
Ovis1.6-Gemma2-9B	0.695	0.555	0.740	0.933	0.639
mPLUG-Owl3	0.640	0.543	0.700	0.900	0.571
claude-3-opus-20240229	0.775	0.542	0.860	0.700	0.763
claude-3-5-sonnet-20240620	0.784	0.540	0.760	0.900	0.769
llava-onevision-qwen2-7b-ov-hf	0.686	0.537	0.633	0.733	0.699
deepseek-vl2-small	0.733	0.532	0.720	0.833	0.718
yi-vision-v2	0.699	0.528	0.780	0.800	0.654
Mantis-8B-Idefics2	0.674	0.516	0.740	0.933	0.603
XinYuan-VL-2B-Instruct	0.648	0.511	0.680	0.867	0.596
VideoLLaMA2-72B	0.716	0.489	0.640	0.800	0.724

Model	**Accuracy**	θ_{total}	θ_{color}	θ_{shape}	θ_{size}
InternVL-Chat-V1-2-Plus	0.691	0.453	0.660	0.800	0.679
LLaVA-Video-72B-Qwen2	0.733	0.451	0.640	0.867	0.737
video_chat_7b	0.458	0.450	0.587	0.655	0.473
grok-2-vision-1212	0.746	0.427	0.620	0.833	0.769
llava-llama-3-8b	0.606	0.411	0.660	0.800	0.551
Qwen2-VL-7B-Instruct	0.699	0.398	0.700	0.933	0.654
MiniCPM-V-2_6	0.661	0.395	0.840	0.867	0.564
MMAlaya2	0.712	0.377	0.660	0.867	0.699
llava-onevision-qwen2-7b-ov-chat-hf	0.606	0.375	0.633	0.733	0.588
LLaVA-NeXT-Video-32B-Qwen	0.720	0.368	0.640	0.867	0.718
Ovis1.6-Llama3.2-3B	0.648	0.343	0.720	0.800	0.596
Llama-3.2V-11B-cot	0.682	0.336	0.640	0.867	0.660
Aria	0.678	0.329	0.800	0.900	0.596
llava-onevision-qwen2-72b-si-hf	0.665	0.326	0.620	0.800	0.662
Phi-4-multimodal-instruct	0.644	0.300	0.780	0.933	0.545
Qwen2-VL-2B-Instruct	0.661	0.298	0.900	0.833	0.551
llava_next_72b	0.648	0.275	0.680	0.800	0.609
Mantis-8B-siglip-llama3	0.623	0.262	0.660	0.833	0.574
Eagle-X4-13B-Plus	0.551	0.246	0.575	0.950	0.633
llava-onevision-qwen2-7b-si-hf	0.576	0.225	0.653	0.700	0.539
InternVL2_5-26B	0.665	0.220	0.580	0.867	0.654
Llama-3-VILA1.5-8B-Fix	0.665	0.192	0.560	0.767	0.679
pllava-7b	0.551	0.187	0.571	0.750	0.540
hunyuan-vision	0.661	0.183	0.700	0.900	0.603
Phi-3.5-Vision	0.665	0.166	0.660	0.667	0.667
Idefics3-8B-Llama3	0.636	0.153	0.720	0.767	0.583
VILA1.5-40B	0.682	0.152	0.560	0.733	0.712
InternVL-Chat-V1-5	0.665	0.147	0.560	0.867	0.660
VideoLLaMA2-7B	0.470	0.145	0.440	0.567	0.462
VideoChat2_HD_stage4_Mistral_7B_hf	0.475	0.144	0.580	0.667	0.412
Video-LLaVA-7B	0.525	0.125	0.700	0.533	0.468
InternVL2_5-38B	0.606	0.124	0.560	0.700	0.603
llava_next_110b	0.648	0.123	0.600	0.633	0.667
qwen-vl-plus-2025-01-25	0.644	0.121	0.720	0.867	0.596
Qwen2.5-VL-7B-Instruct	0.653	0.116	0.760	0.867	0.577
VideoLLaMA2-7B-16F	0.593	0.109	0.580	0.733	0.571
Qwen2.5-VL-7B-Instruct_video	0.644	0.098	0.740	0.867	0.571
internlm-xcomposer2d5-7b	0.331	0.089	0.641	0.818	0.473
InternVL2-40B	0.682	0.080	0.560	0.933	0.673
InternVL2-8B	0.644	0.080	0.540	0.800	0.647
deepseek-vl-7b-chat	0.525	0.078	0.463	0.900	0.626
video_chat_13b	0.419	0.066	0.457	0.607	0.466
InternVL2-76B	0.669	0.060	0.620	0.933	0.635
InternVL2_5-1B	0.555	0.040	0.540	0.633	0.545
emu2-chat	0.530	0.029	0.480	0.700	0.513
LLaVA-Video-7B-Qwen2-Video-Only_multi_frame	0.504	0.028	0.520	0.700	0.462
Pixtral-12B-2409	0.606	0.027	0.600	0.800	0.571
llava_next_llama3	0.449	0.025	0.525	0.750	0.504
InternVL2-26B	0.648	0.015	0.580	0.767	0.647
VILA1.5-3b	0.513	0.012	0.540	0.700	0.468
SmolVLM	0.602	0.007	0.680	0.667	0.564
LLaVA-NeXT-Video-7B	0.441	0.000	0.804	0.556	0.382
gpt-4o-mini	0.627	-0.035	0.460	0.900	0.628
InternVL2-4B	0.657	-0.043	0.500	0.833	0.673
InternVL2_5-8B	0.551	-0.053	0.520	0.767	0.519
Mini-InternVL-Chat-4B-V1-5	0.602	-0.065	0.520	0.833	0.583
Janus-1.3B	0.445	-0.071	0.450	0.400	0.568

Model	Accuracy	θ_{total}	θ_{color}	θ_{shape}	θ_{size}
llava_next_interleave_7b_dpo	0.525	-0.096	0.540	0.733	0.481
deepseek-vl2	0.619	-0.102	0.540	0.767	0.615
deepseek-vl2-tiny	0.606	-0.107	0.460	0.800	0.615
internlm-xcomposer2-7b	0.623	-0.109	0.580	0.828	0.614
deepseek-vl-1.3b-chat	0.500	-0.109	0.375	0.600	0.655
Llama-3.2-90B-Vision-Instruct	0.606	-0.115	0.480	0.500	0.667
InternVL2-2B	0.475	-0.124	0.580	0.800	0.378
Llama-3-VILA1.5-8B	0.534	-0.126	0.440	0.633	0.545
InternVL2_5-4B	0.576	-0.133	0.560	0.724	0.558
Mantis-8B-clip-llama3	0.530	-0.142	0.480	0.700	0.513
LLaVA-NeXT-Video-7B-DPO	0.453	-0.142	0.646	0.556	0.418
llava-onevision-qwen2-0.5b-ov-hf	0.517	-0.142	0.440	0.600	0.526
InternVL-Chat-V1-1	0.589	-0.149	0.480	0.733	0.596
LLaVA-Video-7B-Qwen2-Video-Only	0.462	-0.152	0.540	0.633	0.404
claude-3-sonnet-20240229	0.568	-0.161	0.480	0.800	0.555
llava-onevision-qwen2-0.5b-si-hf	0.466	-0.171	0.420	0.500	0.484
Qwen2.5-VL-3B-Instruct	0.542	-0.205	0.380	0.867	0.532
pllava-13b	0.470	-0.205	0.444	0.556	0.524
Llama-3-LongVILA-8B-128Frames	0.542	-0.216	0.540	0.833	0.487
Qwen2.5-VL-3B-Instruct_video	0.542	-0.220	0.400	0.833	0.532
claude-3-haiku-20240307	0.610	-0.223	0.440	0.733	0.641
llava_next_interleave_7b	0.530	-0.238	0.540	0.533	0.526
Chat-UniVi-13B	0.271	-0.269	0.750	0.625	0.522
VideoLLaMA2.1-7B-16F	0.585	-0.296	0.500	0.600	0.609
qwen_chat	0.508	-0.385	0.380	0.700	0.513
VILA1.5-13b	0.534	-0.449	0.460	0.567	0.551
InternVL2_5-2B	0.415	-0.455	0.380	0.667	0.378
Qwen2.5-Omni-7B	0.547	-0.468	0.340	0.733	0.577
Eagle-X5-34B-Plus	0.445	-0.523	0.225	0.500	0.619
Mini-InternVL-Chat-2B-V1-5	0.407	-0.551	0.380	0.800	0.342
Vintern-3B-beta	0.538	-0.587	0.320	0.533	0.609
Eagle-X4-8B-Plus	0.424	-0.601	0.184	0.750	0.561
idefics_9b_instruct	0.453	-0.677	0.383	0.724	0.500
idefics_80b_instruct	0.525	-0.686	0.347	0.633	0.564
JanusFlow-1.3B	0.347	-0.692	0.300	0.450	0.455
VILA1.5-3b-s2	0.500	-0.702	0.320	0.533	0.551
Llama-3.2-11B-Vision-Instruct	0.517	-0.707	0.340	0.567	0.575
OpenFlamingo-9B-vitl-mpt7b	0.483	-0.725	0.340	0.500	0.526
Janus-Pro-1B	0.369	-0.734	0.244	0.700	0.450
InternVL2-1B	0.453	-0.787	0.340	0.633	0.455
Vintern-1B-v2	0.496	-0.787	0.306	0.600	0.538
Eagle-X5-7B	0.403	-0.862	0.200	0.450	0.561
Eagle-X5-13B	0.398	-0.872	0.200	0.500	0.547
Valley2-7b	0.051	-0.882	0.286	0.500	0.267
video_chatgpt-7B	0.225	-0.926	0.271	0.357	0.210
Chat-UniVi	0.318	-0.950	0.340	0.414	0.297
qwen_base	0.140	-1.023	0.400	0.417	0.558

A.2 ITEM-LEVEL DISCRIMINATION AND DIFFICULTY PARAMETERS ESTIMATED FROM THE 2PL IRT MODEL.

Table 2: a = discrimination, b = difficulty, passrate = proportion correct.

Item ID	Subtype	a	b	Passrate
400488	Size	8.241	0.204	0.503
400462	Color	6.436	-0.327	0.820
400487	Size	6.178	0.238	0.480
400486	Size	5.615	0.172	0.520
400515	Color	5.434	0.056	0.597
400500	Color	5.430	-0.293	0.793
400502	Color	5.338	-0.260	0.779
400444	Size	5.184	0.261	0.464
400449	Color	4.969	-0.263	0.773
400440	Size	4.904	0.189	0.507
400475	Size	4.855	-0.376	0.818
400458	Size	4.848	0.230	0.487
400441	Size	4.846	0.335	0.423
1732	Size	4.793	-0.055	0.656
400459	Size	4.753	-0.445	0.840
400484	Size	4.676	0.124	0.550
400467	Color	4.646	-0.307	0.788
400507	Size	4.598	0.373	0.388
400504	Color	4.552	0.041	0.603
400442	Size	4.539	0.480	0.336
400482	Size	4.485	0.224	0.490
400513	Size	4.417	0.362	0.405
400453	Color	4.263	-0.055	0.651
400499	Size	4.234	0.098	0.563
400472	Size	4.219	0.064	0.587
400517	Color	4.104	-0.295	0.772
400470	Size	4.023	0.242	0.477
400456	Size	3.995	-0.496	0.838
1740	Shape	3.977	0.056	0.583
400509	Size	3.873	0.224	0.490
1731	Size	3.790	0.290	0.446
400501	Color	3.690	0.440	0.369
1745	Size	3.672	0.259	0.467
400446	Size	3.667	0.253	0.477
400468	Color	3.551	0.351	0.423
1762	Size	3.396	0.108	0.550
400523	Color	3.385	0.395	0.397
400438	Size	3.380	0.468	0.360
1861	Size	3.376	-0.072	0.649
1744	Color	3.213	0.036	0.593
400471	Size	3.212	0.907	0.174
1765	Size	3.142	0.448	0.368
1730	Color	3.074	1.063	0.132
400437	Color	3.034	0.523	0.340
1736	Size	2.875	0.478	0.365
400520	Color	2.857	-0.026	0.616
1776	Size	2.830	1.140	0.121
1826	Shape	2.825	-0.724	0.874
1733	Color	2.806	0.283	0.460
400495	Size	2.754	0.324	0.443
1803	Size	2.662	0.562	0.329
1798	Size	2.641	0.323	0.440

Item ID	Subtype	a	b	Passrate
400524	Color	2.622	0.342	0.433
1742	Size	2.616	-0.419	0.772
1812	Size	2.587	0.278	0.461
1852	Size	2.560	0.516	0.351
400477	Size	2.551	0.408	0.404
1882	Color	2.511	-0.629	0.829
1814	Size	2.505	0.687	0.283
400466	Color	2.482	0.591	0.327
1821	Size	2.466	-0.130	0.651
1817	Shape	2.447	-0.177	0.667
1802	Size	2.407	0.334	0.439
400522	Color	2.373	0.497	0.368
1858	Size	2.339	0.776	0.259
1804	Color	2.330	0.452	0.388
400490	Size	2.282	0.225	0.487
1875	Size	2.281	0.260	0.473
1763	Size	2.275	0.405	0.409
1743	Color	2.268	1.018	0.185
1707	Shape	2.177	-0.730	0.834
1800	Size	2.175	0.457	0.385
1250	Shape	2.157	-0.295	0.713
1783	Size	2.104	0.376	0.430
1241	Shape	2.087	-0.339	0.726
1853	Shape	2.040	-0.497	0.763
1756	Color	2.034	0.120	0.533
1784	Size	1.946	-0.146	0.632
1874	Shape	1.908	-1.248	0.914
1210	Size	1.861	0.352	0.444
1747	Size	1.856	0.714	0.309
1248	Size	1.838	0.690	0.321
1813	Shape	1.835	0.252	0.480
1836	Shape	1.786	-0.009	0.576
1223	Size	1.740	0.024	0.574
1240	Size	1.681	-0.143	0.630
1805	Size	1.630	0.711	0.325
400521	Color	1.562	0.993	0.253
400489	Size	1.553	0.901	0.278
400525	Color	1.547	0.999	0.253
1247	Shape	1.512	0.106	0.537
1001208	Size	1.458	-1.811	0.941
400519	Color	1.416	0.807	0.318
1737	Color	1.394	0.893	0.296
1221	Shape	1.378	-0.457	0.699
1901	Color	1.362	-1.141	0.841
1816	Shape	1.357	-0.431	0.682
1208	Size	1.356	-1.368	0.881
1242	Color	1.345	0.272	0.482
1237	Color	1.332	0.495	0.415
1001723	Size	1.310	0.479	0.424
1725	Size	1.216	1.547	0.181
1888	Color	1.200	0.146	0.513
1001724	Color	1.194	-2.524	0.959
1724	Color	1.190	-3.082	0.980
1773	Size	1.174	2.061	0.114
2001723	Size	1.164	0.646	0.388
1894	Size	1.145	0.656	0.382
400436	Size	1.124	1.252	0.250

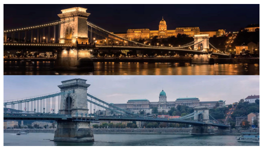
Item ID	**Subtype**	a	b	**Passrate**
1883	Size	1.088	1.091	0.289
1791	Size	1.081	0.515	0.423
1746	Color	1.061	0.945	0.322
1212	Shape	1.052	-2.385	0.934
1224	Color	1.039	-1.650	0.867
1001725	Size	1.009	1.826	0.174
1001704	Shape	0.987	-0.860	0.733
1217	Shape	0.967	-1.744	0.861
1222	Color	0.954	-2.588	0.933
1751	Size	0.923	-2.097	0.887
1220	Size	0.915	-1.449	0.815
1779	Size	0.914	-2.191	0.893
1244	Size	0.878	-2.728	0.927
1831	Shape	0.871	0.333	0.473
3001723	Size	0.870	0.126	0.517
1245	Size	0.859	0.956	0.350
400445	Size	0.852	-2.913	0.932
1704	Shape	0.845	-1.695	0.827
1225	Size	0.844	-2.779	0.925
1249	Size	0.841	-2.301	0.890
1723	Size	0.839	0.470	0.449
1230	Size	0.839	-2.388	0.896
1239	Color	0.823	-2.250	0.881
1228	Color	0.820	-1.720	0.825
1209	Size	0.820	-2.427	0.895
1215	Shape	0.814	-2.115	0.867
1001773	Size	0.812	2.859	0.107
1699	Color	0.807	-2.586	0.901
1227	Size	0.789	-1.469	0.787
400478	Size	0.785	-2.164	0.861
1232	Shape	0.779	-3.015	0.926
1698	Color	0.779	-2.635	0.899
1785	Shape	0.766	-2.463	0.882
1702	Size	0.757	-3.188	0.928
1235	Size	0.750	-1.174	0.735
1891	Color	0.742	0.322	0.477
1238	Shape	0.740	-2.622	0.890
1809	Size	0.736	-2.294	0.860
1739	Size	0.732	-2.738	0.894
1728	Size	0.722	-1.973	0.824
1769	Size	0.704	-1.618	0.778
1706	Shape	0.703	-2.218	0.844
1863	Shape	0.702	-1.711	0.789
1734	Size	0.699	-2.748	0.886
400455	Size	0.698	-1.764	0.793
1219	Shape	0.684	-1.674	0.784
1703	Size	0.682	-2.900	0.893
400457	Size	0.682	-3.629	0.934
1788	Size	0.668	-2.468	0.854
1903	Size	0.667	-3.082	0.900
1758	Size	0.667	-3.099	0.901
1793	Size	0.667	-1.854	0.795
1226	Size	0.664	-2.006	0.812
400448	Size	0.661	-2.561	0.860
1246	Color	0.659	-2.559	0.861
1771	Size	0.656	-0.462	0.605
400480	Size	0.651	-2.878	0.882

Item ID	Subtype	a	b	Passrate
1700	Color	0.642	-3.123	0.895
1229	Color	0.642	-1.718	0.776
1218	Color	0.640	-2.403	0.843
2001773	Size	0.634	3.240	0.127
1701	Color	0.624	-2.788	0.867
400479	Shape	0.620	-3.110	0.887
1708	Size	0.615	-1.528	0.742
2001725	Size	0.615	1.499	0.313
1887	Size	0.604	-3.012	0.875
1810	Size	0.594	-1.291	0.707
1849	Size	0.593	-2.856	0.860
1774	Size	0.593	-3.036	0.873
1782	Size	0.578	-1.556	0.733
1766	Size	0.563	-1.345	0.705
1873	Size	0.553	-0.680	0.619
1792	Size	0.551	-2.343	0.803
1829	Shape	0.538	-2.936	0.847
1868	Size	0.537	-2.531	0.813
1764	Size	0.531	-0.399	0.578
1906	Size	0.525	-2.247	0.783
1825	Shape	0.522	-2.507	0.807
1881	Color	0.513	-2.324	0.787
400483	Size	0.506	-1.149	0.664
400511	Size	0.506	-2.526	0.800
1711	Size	0.495	-1.912	0.740
1815	Size	0.489	-1.442	0.691
1854	Size	0.483	-3.171	0.840
1234	Color	0.483	-1.318	0.679
1786	Size	0.472	-1.392	0.680
1869	Size	0.469	-2.487	0.781
1847	Size	0.465	-2.165	0.753
1741	Size	0.464	-1.165	0.653
1778	Size	0.455	-2.041	0.737
1845	Size	0.454	-1.819	0.716
1860	Shape	0.444	-1.384	0.671
1781	Size	0.440	-0.895	0.618
1801	Size	0.438	-0.880	0.616
1738	Size	0.437	1.258	0.384
1859	Shape	0.433	-1.459	0.673
1844	Size	0.432	-2.180	0.740
1880	Size	0.425	-1.788	0.702
1885	Size	0.422	-1.469	0.671
1895	Size	0.420	-1.209	0.645
1770	Size	0.417	-0.522	0.576
1797	Size	0.410	-0.126	0.533
1871	Size	0.410	-1.349	0.656
1710	Size	0.409	-1.685	0.687
1897	Size	0.402	1.797	0.342
1808	Size	0.396	-0.699	0.589
1777	Size	0.389	-1.002	0.617
1796	Size	0.387	1.847	0.342
400473	Size	0.385	0.570	0.463
1211	Size	0.379	-2.272	0.726
1787	Size	0.376	0.166	0.503
1780	Size	0.372	-0.227	0.539
1748	Size	0.366	-2.779	0.755
1902	Size	0.365	-0.394	0.555

Item ID	**Subtype**	a	b	**Passrate**
1856	Size	0.365	-0.529	0.567
1790	Size	0.365	-0.813	0.592
1729	Size	0.363	1.273	0.401
400493	Size	0.362	-2.697	0.747
1806	Size	0.357	0.573	0.466
1840	Size	0.351	0.001	0.517
1705	Size	0.319	-1.680	0.651
400496	Size	0.300	-2.999	0.733
1752	Size	0.300	-3.221	0.747
400460	Size	0.297	-2.894	0.725
400461	Size	0.293	-3.380	0.752
400447	Color	0.248	-2.001	0.642
400498	Size	0.239	-1.672	0.617

A.3 Additional Examples

A



In the picture, has the actual color of the bridge itself changed? Please answer with Yes or No and provide an explanation.

Correct Answer: No

GPT4o: No. The actual color of the bridge itself has not changed the difference in appearance between the two photos is primarily due to the lighting conditions the top image is taken at night with artificial lighting illuminating the bridge and surrounding area, giving the bridge a warm, glowing appearance the bottom image is taken during the day with natural daylight, which shows the bridge in its true color without the effects of artificial lighting.

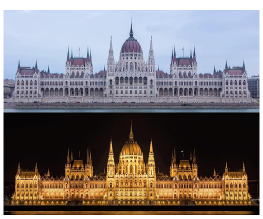
B

In the picture, has the actual color of the building itself changed? Please answer with Yes or No and provide an explanation.

Correct Answer: No

GPT4o: No. The actual color of the building itself has not changed the difference in appearance is due to the time of day and the lighting conditions in the top image, the building is shown during daylight hours, which provides natural lighting in the bottom image, the building is illuminated at night with artificial lights, which enhance and highlight certain architectural features, giving it a different visual effect the building's actual color remains the same; it is the lighting conditions that create the variation in appearance.

C



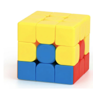
In the picture, has the actual color of the building itself changed? Please answer with Yes or No and provide an explanation.

Correct Answer: No

GPT4o: No. The actual color of the building itself has not changed the two images show the same building at different times of day and under different lighting conditions in the top image, the building is shown during daylight, which presents its natural colors in the bottom image, the building is illuminated at night with artificial lighting, which gives it a warmer, golden appearance the change in color is due to the lighting, not a change in the building's actual colors

Figure 5: **Additional Sample Tasks: Color Constancy**.

A



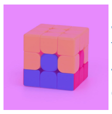
What color is the middle square of the Rubik's Cube? Please select from the following choices:

A: White,
B: Yellow,
C: Blue

Correct Answer: C

GPT4o: C blue. In the image provided, the rubik's cube's middle square on the visible face is blue this is the face with predominantly yellow stickers, and the middle square is distinctly blue

B



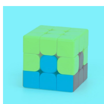
What color is the middle square of the Rubik's Cube? Please select from the following choices:

A: Purple,
B: Yellow,
C: Blue

Correct Answer: C

GPT4o: C blue. The central square in the front face of the rubik's cube, which is visible in the image, is blue

C



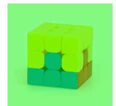
What color is the middle square of the Rubik's Cube? Please select from the following choices:

A: Red,
B: Yellow,
C: Blue

Correct Answer: C

GPT4o: C blue. The image provided shows a rubik's cube with several visible faces the middle square on the front face of the cube is distinctly blue

D



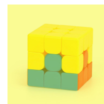
What color is the middle square of the Rubik's Cube? Please select from the following choices:

A: Green,
B: Yellow,
C: Blue

Correct Answer: C

GPT4o: A green. The middle square of the rubik's cube shown in the image is color green this can be determined by observing the central square on the visible face of the cube

E



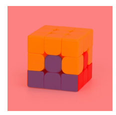
What color is the middle square of the Rubik's Cube? Please select from the following choices:

A: Red,
B: Yellow,
C: Blue

Correct Answer: C

GPT4o: C Blue. The middle square on the visible side of the rubik's cube in the image is blue

F



What color is the middle square of the Rubik's Cube? Please select from the following choices:

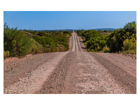
A: Red,
B: Yellow,
C: Blue

Correct Answer: C

GPT4o: None. The color of the middle square of the rubik's cube is purple thus, none of the given options (a red, b yellow, c blue) are correct. The middle square of the visible face on the rubik's cube is purple

Figure 6: **Additional Sample Tasks: Color Constancy**.

A

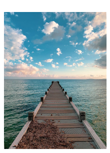


In the picture, does the actual width of the road get narrower from near to far? Please answer with Yes or No and provide an explanation.

Correct Answer: No

GPT4o: No. the actual width of the road does not get narrower from near to far the road appears to get narrower due to the effect of perspective, which causes objects to seem smaller as they get further away from the viewer this visual effect is typical in photographs or any scenes with a defined vanishing point the road itself remains the same width throughout

B

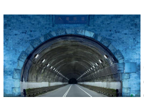


In the picture, do the actual sizes of the wooden posts on both sides of the pier get smaller from near to far? Please answer with Yes or No and provide an explanation.

Correct Answer: No

GPT4o: No, the actual sizes of the wooden posts on both sides of the pier do not get smaller from near to far in the picture, it only appears that the posts are getting smaller due to the perspective effect as objects go farther away from the viewer, they appear smaller even if their actual size remains unchanged this is a common optical illusion in photography and real-life views, where parallel lines seem to converge at a vanishing point in the distance

C

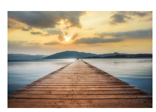


Is this tunnel getting narrower? Please answer with Yes or No and provide an explanation.

Correct Answer: No

GPT4o: No, this tunnel is not getting narrower the illusion is created by the perspective of the photograph as the tunnel extends into the distance, objects further away from the camera appear smaller due to the rules of linear perspective this makes the far end of the tunnel appear narrower than the entrance even though the tunnel's actual dimensions do not change the effect is enhanced by the parallel lines of lights and the road markings converging towards the vanishing point

D

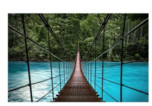


In the picture, is the actual width of the pier consistent from near to far? Please answer with Yes or No and provide an explanation.

Correct Answer: Yes

GPT4o: No. The actual width of the pier remains consistent from near to far the apparent decrease in width is due to perspective as objects extend into the distance, they appear smaller and narrower, creating an illusion that they are converging this is a common visual effect in photography and human vision called perspective foreshortening the physical width of the pier itself does not change, but our perception of it does

E



Has the actual width of the suspension bridge deck in the picture changed?

Correct Answer: No

GPT4o: No. The actual width of the suspension bridge deck has not changed the bridge appears narrower as it extends into the distance due to the effect of perspective in the photograph perspective causes objects that are farther away to appear smaller than those that are closer to the viewer, creating an illusion that the width is decreasing as you look further along the bridge

F

In the picture, does the actual width of the train tracks get narrower from near to far? Please answer with Yes or No and provide an explanation.

Correct Answer: No

GPT4o: No. The actual width of the train tracks does not get narrower from near to far the apparent narrowing of the train tracks in the image is an example of perspective, a visual phenomenon where parallel lines appear to converge as they recede into the distance this effect occurs because of the way three-dimensional objects are projected onto a two-dimensional surface (the image), creating the illusion that train tracks get closer together as they extend further away from the viewer

Figure 7: **Additional Sample Tasks: Size Constancy**.

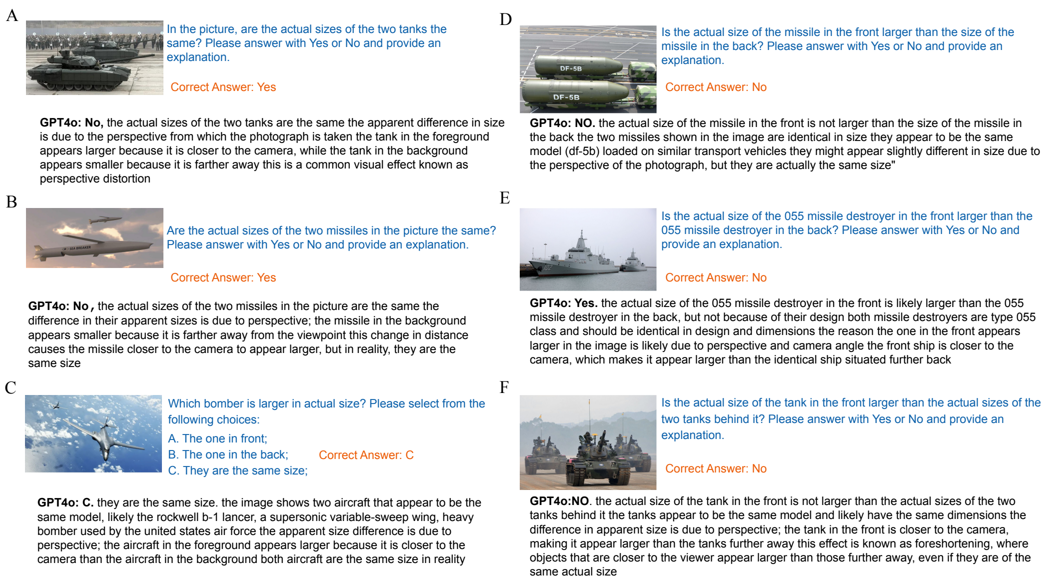


Figure 8: **Additional Sample Tasks: Size Constancy**.

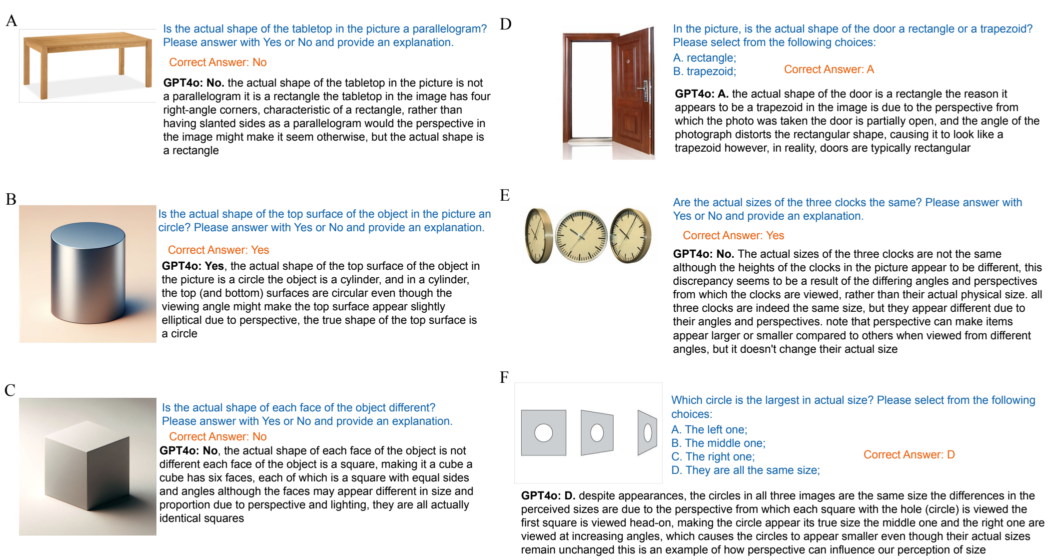


Figure 9: **Additional Sample Tasks: Shape Constancy**.