

Research And Analysis of Artificial Intelligence in Abnormal Image Recognition

Runze Tao *

School of Computer Science and Informatics, University of Liverpool, Liverpool, the United Kingdom

* Corresponding Author Email: RunzeTao23@student.xjtlu.edu.cn

Abstract. Currently, more and more image recognition technologies are being used in AI, and these AIs are gradually being applied in productivity scenarios. This article talks about the performance of the two mainstream AI ChatGPT5 and Qwen VL-Max in recognizing abnormal images. This research divides the abnormal images into four categories: Anatomical Anomalies, Physical Law Violations, Functional and Contextual Incongruities, Scale and Proportion Paradoxes. It is hoped that through this research, we can discover whether different AIs have varying perceptions of abnormal images generated by divergent models, in order to determine how differences in the training dataset impact the model's performance. Through three progressively in-depth questions for each image, researchers found that in most questions, the difference between the two was not significant, and both could identify the issues in the images. However, regarding anatomical anomalies that most mainstream models struggled to detect, ChatGPT, although providing incorrect answers, reflected on them and proactively requested a comparison with normal structural models. Researchers hope that this result will provide insights for developers.

Keywords: Computer vision, Artificial Intelligence model, Image recognition.

1. Introduction

In recent years, with the development of increasingly powerful deep learning architectures and large-scale datasets, there is a big evolution in computer vision and multimodal learning. The breakthrough of convolutional neural networks, exemplified by the success of AlexNet on the ImageNet benchmark marked a turning point by demonstrating the effectiveness of deep layered structures and ReLU activation in large-scale visual recognition tasks [1]. This work not only popularized the use of GPUs for deep learning but also set a new direction for the field by showing that feature learning could substantially outperform hand-crafted visual features. Subsequent innovations further addressed fundamental challenges in scaling deep models—for instance, residual networks (ResNet), introduced by He et al., effectively mitigated the degradation problem in very deep architectures through skip connections, enabling the training of networks with previously unattainable depth and significantly improving accuracy across numerous vision tasks [2]. What's more, there are Vision Transformer (ViT) which can achieve competitive or even superior performance in image classification and CLIP with enabled models to align images and natural language at scale, providing transferable representations for zero-shot recognition [3, 4].

This rapid advancing makes more and more general-purpose large models start Image recognition work. So, this article will discuss about the difference between Qwen from Ali and ChatGPT from OpenAI in Image recognition. hoping to understand through this research whether AI trained on different data will make same errors in certain error-prone recognizable images.

2. Experimental Setup

2.1. Model Selection

The two models selected for this experiment are ChatGPT5 and Qwen VL-Max, chosen for their distinct linguistic training backgrounds—the former being notably trained primarily on English text and the latter on Chinese. Table 1 shows their features:

Table 1. Model Feature

Feature	ChatGPT-5	Qwen VL-Max
Training Data Emphasis	Multilingual balance, primarily English	Chinese-enhanced, optimized for both Chinese and English
Visual Processing Strength	Cross-modal generation (e.g., image-text creation)	High-resolution image parsing, text extraction and localization
Reasoning Mechanism	Dynamic routing + hierarchical reasoning	Three-stage training + vision-language adapter
Typical Application Scenarios	Programming assistance, long-document analysis, personalized interaction	Chinese document understanding, chart analysis, industrial inspection

2.2. Test Dataset Construction

To systematically evaluate the models' comprehensive ability in recognizing unconventional images, a dedicated test dataset was constructed. This dataset includes four different types of unconventional images to avoid them being too simple referring to MMMU-Pro [5]. The design aims to examine the models' performance beyond their "comfort zone" through targeted challenge categories to deeply explore their capability boundaries.

This test set is specifically designed to stress-test the models' visual grounding and logical reasoning abilities through four typical violation categories:

Anatomical Anomalies (N = 1): Depicts biologically impossible images, such as a hand with only four fingers. This category tests the model's understanding of biological structures and anatomical correctness.

Physical Law Violations (N = 1): Presents physically impossible scenes, such as a piano suspended in mid-air. This category evaluates the model's understanding of physical laws and causal relationships.

Functional and Contextual Incongruities (N = 1): Shows objects in unreasonable contexts, such as a toaster in a forest. This category examines the model's semantic understanding and contextual reasoning abilities.

Scale and Proportion Paradoxes (N = 1): Contains images with impossible size relationships, such as a miniature version of an elephant that can be held in hand. This category tests the model's perception of relative size and proportional relationships.

These images were primarily generated using Stable Diffusion through carefully designed prompts to ensure clear violation of everyday cognitive expectations.

Prompt Design:

To ensure fairness and comprehensiveness in evaluation, the paper used a unified set of fixed prompts for all models. This research reference TextCaps. There is no description about the pictures and ask the model to conclude the pictures [6]. The prompts were specifically designed to elicit responses at different complexity levels in table 2:

Table 2. Question Level

Level 1 (Description)	Please describe this image in detail.
Level 2 (Detailed Q&A)	How many fingers are on the hand in the image? [For anatomical anomalies]
Level 3 (Reasoning)	Is there anything unusual or illogical in this image? If so, please explain.

This multi-level prompt strategy enables systematic evaluation of the models' basic perception and higher-order reasoning capabilities across different types of violations.

2.3. Evaluation Metrics

To quantitatively compare the overall performance of ChatGPT and Qwen, the paper designed a specialized comprehensive scoring system. Each model's responses in each image category are evaluated and assigned corresponding scores. The final composite score employs a weighted average calculation method, ensuring balanced consideration of detection capabilities for various types of violations.

Scoring Criteria: Unconventional Images (25% weight for each subcategory): Scoring for each unconventional image subcategory focuses on the model's ability to detect and reason about specific types of violations: Correct identification and reasonable explanation of the violation: 1 point; Detection of anomaly but insufficient or inaccurate explanation: 0.5 points; Failure to identify the anomaly or production of hallucinatory explanations: 0 points; The final score for each subcategory is the subcategory score multiplied by the weight coefficient (0.25). **Final Score Calculation :** The final composite score (S) for each model is calculated using the following formula:

$$S = (\text{Anatomical Anomaly Score} \times 0.25) + (\text{Physical Violation Score} \times 0.25) + (\text{Contextual Incongruity Score} \times 0.25) + (\text{Scale Paradox Score} \times 0.25)$$

This score ranges from 0 to 1 (or 0% to 100%), providing a standardized metric for overall performance comparison while maintaining sensitivity to various types of reasoning capabilities.

3. Results and Analysis

3.1. Quantitative Results

According to the quantitative evaluation results, ChatGPT-5 slightly outperformed Qwen VL-Max with an overall score of 0.875 compared to 0.75. This difference reflects the models' capabilities in judging the logical consistency of image content. The following analysis is conducted across four dimensions:

In detecting anatomical anomalies, both models correctly identified that the hand had five fingers. However, Qwen VL-Max demonstrated logical deviations in its subsequent inconsistency analysis. It spent significant effort discussing secondary details such as skin texture and lighting effects while overlooking the most critical anatomical anomaly—the clearly disproportionate finger lengths. In contrast, ChatGPT-5 directly pointed out key issues such as the abnormally long pinky finger and the unnatural positioning of the thumb, demonstrating more precise anomaly detection capabilities.

Regarding violations of physical laws, both models successfully identified the primary anomaly of the floating piano. Yet, Qwen VL-Max again engaged in excessive description, delving into secondary issues such as whether the shadow details perfectly matched. ChatGPT-5 remained focused on the core contradiction: the conflict between the piano's weight and the principles of buoyancy, as well as the lack of a plausible support mechanism in the street scene. This focused analysis made its response more logically convincing.

In assessing functional and contextual incongruities, both models accurately recognized that the toaster in the forest had a power source issue. However, Qwen VL-Max's analysis contained logical flaws—while acknowledging that there was "no visible power source," it simultaneously speculated that it might be "connected to an invisible power source." This self-contradiction undermined the credibility of its response. ChatGPT-5, on the other hand, maintained logical consistency by clearly highlighting the fundamental contradiction between the absence of a power source and the device's operational state, without engaging in over-speculation.

Finally, in analyzing scale and proportion paradoxes, Qwen VL-Max provided a detailed description of the miniature elephant's morphological features but overlooked the most critical reference point—the standard size of a human hand. This omission weakened the substantiation of its analysis. ChatGPT-5, however, explicitly estimated the size as "5–7 cm" and strengthened its argument by comparing it to the size of a real baby elephant (approximately 90 kg), resulting in a more persuasive proportional analysis.

In summary, ChatGPT-5’s advantages manifest in three aspects: first, a stronger ability to capture core contradictions, enabling it to quickly bypass secondary details and focus on key issues; second, better logical consistency, avoiding self-contradictory statements; and third, more precise use of reference systems, providing quantitative analytical support. Although Qwen VL-Max performed well in detailed front-end descriptions, it tended to deviate from the main points or contradict itself in subsequent logical reasoning, which may be the main reason for its slightly lower score. This difference reflects subtle disparities in the two models’ visual-language integration capabilities, particularly in maintaining logical consistency. Here is the final score of the two models in table 3:

Table 3. Score of Qwen VL-max and ChatGPT 5

	Anatomical Anomalies	Physical Law Violations	Functional and Contextual Incongruities	Scale and Proportion Paradoxes	Total
Qwen VL-max	0	1	1	1	0.75
ChatGPT 5	0.5	1	1	1	0.875

3.2. Qualitative Analysis

Overall, both ChatGPT and Qwen VL-Max demonstrated strong multimodal comprehension capabilities when confronted with unconventional images. Their final composite scores (ChatGPT: 0.875; Qwen VL-Max: 0.75) indicate relatively close overall performance. However, a detailed analysis of their responses, particularly their performance in the "Anatomical Anomalies" category, reveals fundamental differences in their underlying reasoning processes. This result is consistent with the results of Fu et al. using the MME test [7].

3.2.1 Commonality in Core Flaw: A Priori Preference for "Normality"

When processing the image of the hand with four fingers, both models initially defaulted to the strong prior knowledge that "a human hand must have five fingers." ChatGPT directly provided the incorrect answer of "five" when asked "how many fingers"; Qwen completely failed to recognize the numerical anomaly, and all subsequent analysis was based on the erroneous premise of "this is a five-fingered hand." This suggests that when current large multimodal models encounter visual evidence that contradicts strong prior knowledge, visual perception can be overridden by cognitive bias, causing them to "see" what they believe should be there rather than what is actually present. This is a common and concerning flaw in such models.

3.2.2 Divergence in Reasoning Process: Active Reasoning vs. Static Description

Despite both failing the initial count, the subsequent behavior of the two models was markedly different, highlighting a key distinction in their capabilities:

After providing the initial incorrect answer, ChatGPT's reasoning process displayed an active contradiction detection mechanism. It detected the "unnaturalness" of the image (e.g., disproportionate fingers, awkward thumb position, ring appearing pasted on) and attempted to use these secondary, yet actual, visual clues to argue the "illogical" nature of the image. More importantly, it proactively offered to use a sketch overlay to pinpoint the issues. This behavior suggests that ChatGPT's model architecture incorporates a stronger tendency for analytical reasoning and interactive verification. Even starting from a wrong premise, its reasoning path still sought evidence to support a conclusion.

Qwen's performance resembled that of an advanced "image describer" rather than an "analyst." It fully accepted the five-finger premise and dedicated all its computational and reasoning power to generating a highly detailed but misdirected static description [8]. It focused on secondary details like the ring's design, skin texture, and background rendering style, citing these as the primary reasons for being "illogical." This indicates that Qwen is slightly inferior in critical visual reasoning and verifying its initial perceptions. It tends to accept the surface structure of the visual input and expand upon it, rather than re-evaluating the basic facts of the input. Perhaps in the future, windowed attention could

give it a greater advantage in response speed [9]. At the same time, this study and the research by Wu et al. found that the text comprehension of visual models is somewhat weaker than that of pure text models [10].

4. Conclusion

This experiment demonstrates that relying solely on final task scores is insufficient for a comprehensive assessment of multimodal models. A nuanced analysis of their failure cases can better reveal the models' internal mechanisms.

GPT's strength lies in its analytical and interactive nature. It tends to treat images as objects needing verification and dissection, exhibiting a "detective"-like reasoning mode that is crucial for solving complex and counter-intuitive problems.

Qwen's strength lies in its descriptive thoroughness and generative fluency. It may hold an advantage in understanding and describing conventional scenes and tasks requiring cultural context (e.g., abstract art), but its reasoning relies more heavily on the model's inherent prior knowledge, making it more susceptible to being solidified by "hallucinations." However, there is Naive Dynamic Resolution mechanism in Qwen2 - VL. This technology can make the model be like human and have an ability to beyond ChatGPT 4O.

This research underscores that future model evaluations should focus more on their reasoning processes and behavioral patterns, not just the accuracy of their final output. Furthermore, it points the way for model improvement: reducing the negative impact of pre-existing knowledge and enhancing the model's ability to actively verify visual evidence and analyze contradictions are key to improving their reliability.

If you follow the “checklist” your paper will conform to the requirements of the publisher and facilitate a problem-free publication process.

References

- [1] Krizhevsky A, Ilya S, and Geoffrey E. H. ImageNet classification with deep convolutional neural networks. *Communications of the ACM* 60, no. 6 2017: 84–90.
- [2] He K M, Zhang X Y, Ren S Q, and Sun J. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770-778. 2016.
- [3] Dosovitskiy A, Lucas B, Alexander K, Dirk W, Zhai X H, Thomas U, Mostafa D et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [4] Radford A, Jong W K, Chris H, Aditya R, Gabriel G, Sandhini A, Girish S et al. Learning Transferable Visual Models from Natural Language Supervision. In *International Conference on Machine Learning*, pp. 8748-8763. PMLR, 2021.
- [5] Xiang Y, Zheng T Y, Ni Y S, Wang Y B, et al. MMMU-Pro: A More Robust Multi-Discipline Multimodal Understanding Benchmark. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.
- [6] Yang W, Wang S L, Yang H, et al. An Early Evaluation of GPT-4V(ision). *arXiv preprint arXiv:2310.16534*, 2023.
- [7] Fu C Y, Chen P X, Shen Y H, et al. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. *arXiv preprint arXiv:2306.13394*, 2023.
- [8] Peng W, Bai S, Tan S N, et al. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191*, 2024.
- [9] Shuai B, Chen K Q, Liu X J, et al. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025.
- [10] Sidorov O, Hu R H, Rohrbach M, and Singh A. TextCaps: a Dataset for Image Captioning with Reading Comprehension. *arXiv preprint arXiv:2003.12462*, 2020.