Proceedings of CONF-SEML 2026 Symposium: Computational Analysis and Modeling in Complex Intelligent Systems
DOI: 10.54254/2755-2721/2026.GU31929

Evaluating Reliability and Error Structure in Image-Based AI Model Outputs

Fanze Meng

College of Liberal Arts, University of Minnesota, Minneapolis, USA
meng0282@umn.edu

Abstract. Image-based artificial intelligence models are widely applied in data science tasks such as image classification, object recognition, and visual content generation. In practice, model outputs are often regarded as reliable once acceptable accuracy levels are achieved on benchmark datasets. However, empirical evidence shows that image-based AI systems frequently exhibit structured and non-random error patterns. In image generation tasks, errors commonly arise from an overreliance on statistical correlations learned from training data, limited semantic grounding, and weak constraints on physical and contextual consistency. These limitations can lead to outputs that appear visually coherent while containing incorrect or non-existent objects, implausible spatial relationships, or violations of basic visual logic. From a data science perspective, such errors are often underexamined because evaluation practices rely heavily on aggregate accuracy metrics and benchmark performance, which tend to obscure localized error concentration and output variability. This study conducts a structured analysis of error patterns and output limitations in image-based AI systems by examining misclassification behavior, generation inconsistencies, and evaluation blind spots observed under realistic data conditions. The findings indicate that understanding AI image generation errors requires focusing on error structure and underlying generation mechanisms rather than relying solely on summary performance measures.

Keywords: Image-based AI models, error rate analysis, output limitations, model evaluation, data science applications

1. Introduction

Artificial intelligence systems designed for image analysis and image generation are widely deployed across applications, such as image classification, object detection, medical imaging, and synthetic image creation. In many real-world workflows, the outputs of these systems are directly used for further analysis or decision-making. Once a model demonstrates strong performance on benchmark datasets, its predictions are often implicitly treated as accurate representations of visual information. Nevertheless, evidence from practical deployment suggests that this assumption does not consistently hold under realistic data conditions.

However, practical deployment has shown that AI systems frequently make errors when analyzing or generating images. These errors commonly include confusion between visually similar

DOI: 10.54254/2755-2721/2026.GU31929

objects, incorrect interpretation of image context, and the generation of images containing implausible or incorrect visual elements. Importantly, such failures tend to occur repeatedly under similar conditions, indicating that they are systematic rather than random.

A key contributing factor to these errors persisting is that image-based AI systems rely heavily on statistical patterns learned from training data. When visual inputs are ambiguous, uncommon, or slightly different from the training distribution, error rates increase in consistent and measurable ways. In image generation tasks, the lack of explicit constraints on semantic meaning and physical consistency further contributes to outputs that appear plausible but are factually or logically incorrect.

This paper focuses on analyzing errors in AI image analysis and generation using data science methods. By analyzing error rates, output variability, and distribution-level behavior, this study aims to identify common error structures and limitations in AI image outputs. Rather than improving model architectures, the goal is to better understand how and why these systems fail when processing visual information.

2. Image-based AI errors in visual analysis and generation

In image-based artificial intelligence systems, errors are most prominently observed during two core tasks: image analysis and image generation. From a data science perspective, these errors are not random prediction failures but reflect systematic limitations in how models learn visual representations from data. When visual inputs are ambiguous, weakly constrained by semantic meaning, or unevenly represented in training distributions, error rates increase in consistent and measurable ways.

This section focuses on two dominant error types that characterize failures in AI-based visual tasks. The first is misclassification in image analysis, where models fail to reliably distinguish visually or semantically similar objects. The second involves semantically inconsistent outputs in image generation, where models produce images that appear visually coherent but contain structural or semantic errors. These two error types represent recurring and fundamental limitations in current image-based AI systems and form the basis of the following analysis.

2.1. Misclassification in image analysis tasks

Misclassification remains a persistent error in image analysis tasks, even for modern vision models trained on large-scale datasets. Empirical studies have shown that models often perform well on benchmark evaluations while exhibiting concentrated errors in fine-grained or visually ambiguous categories. Dosovitskiy et al. demonstrate that vision transformer models achieve strong overall accuracy yet remain sensitive to subtle visual differences when class boundaries are not clearly separable, leading to elevated error rates in ambiguous cases [1].

From an experimental evaluation perspective, robustness studies further reveal that misclassification errors are strongly influenced by distributional shifts. Hendrycks and Dietterich systematically evaluated image classifiers under natural corruptions and found that performance degradation is uneven across visual conditions, indicating that error rates cluster under specific input transformations rather than increasing uniformly [2]. This suggests that misclassification behavior is strongly shaped by how training data represent visual variability.

In addition, representation learning plays a critical role in misclassification patterns. Radford et al. showed that models trained with multimodal contrastive objectives improve generalization but still exhibit category-level confusion when visual semantics overlap [3]. These findings indicate that

DOI: 10.54254/2755-2721/2026.GU31929

misclassification errors persist across model architectures and training paradigms, reflecting structural challenges in visual representation rather than isolated model weaknesses.

2.2. Hallucinated visual features in image generation

In image generation tasks, AI systems exhibit a distinct class of errors characterized by semantically inconsistent outputs. These errors occur when generated images appear visually realistic but contain incorrect object attributes, implausible spatial relationships, or scene-level inconsistencies. Experimental studies on generative diffusion models show that such failures arise because training objectives prioritize distributional similarity over explicit semantic constraints.

Liu et al. conducted a systematic evaluation of text-to-image models and found that object count errors and attribute mismatches frequently occur under compositional or underspecified prompts, even when visual quality remains high [4]. These findings demonstrate that semantic errors are not random artifacts but emerge predictably under specific input conditions.

From the evaluation experience, these generation errors are difficult to detect because standard metrics emphasize perceptual realism rather than semantic correctness. As a result, models may produce outputs that satisfy surface-level visual criteria while violating underlying logical or contextual consistency. This highlights a fundamental limitation in current image generation approaches, where visual plausibility does not guarantee semantic validity.

3. Limitations of current evaluation methods in image-based AI

The error patterns discussed in Section 2 raise a critical question: why do such systematic failures persist despite extensive model evaluation? A key reason lies in the limitations of current evaluation methods commonly used in image-based artificial intelligence. In many data-driven workflows, model performance is assessed using aggregate metrics such as accuracy, mean error rate, or perceptual similarity scores. While these measures are convenient for comparison, they provide limited insight into how errors are distributed across inputs and conditions.

Recent evaluation research indicates that aggregate metrics implicitly assume error homogeneity, an assumption that rarely holds in practice. A 2024 study published in Frontiers in Artificial Intelligence demonstrates that image classification models often exhibit highly uneven uncertainty and error distributions across classes and visual conditions, even when overall accuracy remains stable [5]. The study shows that standard evaluation protocols fail to capture regions of elevated risk where prediction confidence and correctness diverge significantly, particularly in visually ambiguous cases. This observation helps explain why misclassification errors identified in image analysis tasks are frequently underestimated during evaluation.

These limitations are even more pronounced in image generation tasks. Current evaluation methods for generative models tend to prioritize visual fidelity and perceptual realism while overlooking semantic correctness. A large-scale 2024 analysis of text-to-image models reports that generated images can receive high perceptual quality scores while containing incorrect object counts, implausible spatial relationships, or contextually inconsistent elements [6]. Because commonly used metrics are not designed to measure semantic validity, these errors remain largely invisible during standard assessment.

Recent work on trustworthy evaluation further highlights a broader methodological gap. A 2025 study on the evaluation of generative AI models argues that without uncertainty-aware and distribution-sensitive evaluation frameworks, performance comparisons are inherently incomplete

DOI: 10.54254/2755-2721/2026.GU31929

[7]. The study demonstrates that models that appear statistically similar under aggregate metrics may exhibit substantially different error behaviors when analyzed at the output level.

From a data science perspective, these findings indicate that existing evaluation methods are poorly aligned with the types of errors observed in image-based AI systems. While aggregate metrics remain useful for high-level benchmarking, reliance on such metrics obscures structured error patterns and provides limited insight into model behavior. Addressing the failures identified in Section 2 requires evaluation strategies that explicitly analyze error concentration, uncertainty, and semantic consistency rather than relying solely on summary performance measures.

4. Reliability and uncertainty in image-based AI outputs

Building on the systematic error patterns identified in Section 2 and the evaluation limitations discussed in Section 3, this section examines the reliability of image-based AI systems from a data science perspective. In this context, reliability refers not only to average predictive performance, but more importantly to the consistency, stability, and trustworthiness of model outputs across varying input conditions.

Unreliable outputs pose a significant risk because they can propagate hidden errors into downstream analysis, visualization, and decision-making processes. In particular, reliability is closely linked to how uncertainty is expressed by models and how sensitive outputs are to small changes in input data. To maintain structural consistency with earlier sections, this discussion is organized around two task categories: image analysis and image generation.

4.1. Reliability issues in image analysis

In image analysis tasks, reliability extends beyond classification accuracy to include whether predictive confidence or uncertainty estimates meaningfully reflect actual error likelihood. A reliable classifier should not only predict correct labels but also provide confidence signals that indicate when predictions are likely to be unreliable. However, extensive evidence suggests that this alignment is often weak in modern deep learning systems.

A comprehensive survey published in Artificial Intelligence Review systematically examines uncertainty estimation methods in deep neural networks and reports that standard image classifiers frequently produce overconfident predictions even when error probability is high [8]. The survey highlights that confidence calibration degrades significantly under distributional shift, class imbalance, or visually ambiguous inputs—conditions that commonly arise in real-world applications. As a result, uncertainty estimates often fail to function as reliable indicators of prediction risk.

This confidence–error mismatch has direct implications for output reliability and downstream analysis. When model outputs appear confident despite a high likelihood of misclassification, downstream analytical pipelines may treat unreliable predictions as trustworthy data points. Such behavior can introduce systematic bias into subsequent analyses, particularly when errors concentrate in specific subpopulations or visual conditions. The survey further emphasizes that misclassification errors and uncertainty are not evenly distributed across the input space, but instead cluster in localized regions that aggregate metrics fail to expose.

These findings indicate that reliability in image analysis cannot be inferred from average accuracy or confidence scores alone. Instead, reliability assessment requires explicit analysis of uncertainty calibration and error concentration across different input regimes. The empirical patterns

DOI: 10.54254/2755-2721/2026.GU31929

summarized in Table 1 illustrate how confidence–error misalignment undermines the dependable use of image classification outputs.

Table 1. The empirical patterns

Dimension	Observed Behavior	Reliability Implication	Supporting Evidence
Prediction confidence	Models frequently produce overconfident predictions even when error probability is high.	Confidence scores fail to reliably indicate true prediction risk.	[8]
Uncertainty calibration	Confidence calibration degrades under distributional shift, class imbalance, and visually ambiguous inputs.	Reliability varies substantially across input conditions.	[8]
Error distribution	Misclassification errors and uncertainty cluster in localized regions of the input space.	Aggregate accuracy masks concentrated failure regions.	[5,8]
Evaluation visibility	Aggregate performance metrics fail to expose confidence–error misalignment.	Structured reliability risks remain undetected during standard evaluation	[5,8]
Downstream impact	Confident but incorrect predictions are treated as trustworthy data points.	Systematic bias propagates into downstream analytical pipelines.	[5,8]

4.2. Reliability in image generation

Reliability challenges are even more pronounced in image generation tasks, where output stability and semantic consistency are critical. Unlike image analysis, image generation produces high-dimensional visual outputs that are inherently sensitive to input variation. Reliable generation requires that similar inputs yield consistent and semantically coherent outputs.

A recent survey published in Computational Visual Media analyzes a wide range of generative models used for personalized image generation [9]. The study reports that many generative systems struggle to maintain output consistency across closely related prompts or user conditions. Even when the intended semantic content remains unchanged, small input perturbations can result in noticeably different generated images. This behavior indicates limitations in the robustness and reliability of current image generation models.

Such output instability undermines the assumption that generated images can be treated as reproducible data points. When similar inputs lead to divergent outputs, uncertainty becomes difficult to quantify and control using standard evaluation metrics. Moreover, commonly used metrics emphasize visual realism rather than semantic correctness, allowing unreliable outputs to appear acceptable during evaluation.

Beyond academic studies, reliability issues in image generation have also been highlighted in real-world contexts. A 2024 Nature News report documents cases in which AI-generated images appear visually convincing but contain factual inaccuracies or misleading visual elements, raising concerns about their use in scientific communication and data-driven decision-making [10]. These observations suggest that reliability limitations in image generation are not confined to experimental settings, but reflect broader constraints in how current generative models encode and reproduce visual semantics.

5. Conclusion

The above analysis shows that reliability in image-based artificial intelligence systems cannot be adequately characterized by aggregate performance metrics alone. By analyzing both image analysis and image generation tasks, the findings demonstrate that errors in AI visual outputs are often structured, systematic, and closely tied to data distribution characteristics rather than random

DOI: 10.54254/2755-2721/2026.GU31929

prediction noise. Misclassification in image analysis and semantically inconsistent outputs in image generation reflect systematic limitations in how models learn and represent visual information.

The analysis further shows that commonly used evaluation practices, which emphasize aggregate accuracy or perceptual quality, are insufficient to reveal these error patterns. Such metrics tend to obscure localized error concentration, output variability, and uncertainty misalignment, leading to an overestimation of model reliability in real-world applications. As discussed throughout this study, reliability cannot be adequately assessed without examining how confidence, uncertainty, and output stability behave across different input conditions.

Rather than proposing new model architectures, this work highlights the importance of rethinking evaluation and analysis strategies in data science workflows. Understanding AI image errors requires attention to error structure, uncertainty behavior, and reliability constraints at the output level. These insights are essential for the responsible and informed use of image-based AI systems in data-driven analysis, decision-making, and visual content generation.

References

[1] Dosovitskiy A., Beyer L., Kolesnikov A., et al. (2021). An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale. ICLR. https: //arxiv.org/pdf/2010.11929

[2] Hendrycks D., & Dietterich T. (2019). Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. ICLR. https: //arxiv.org/abs/1903.12261

[3] Radford A., Kim J. W., Hallacy C., et al. (2021). Learning Transferable Visual Models from Natural Language Supervision. ICML. https: //proceedings.mlr.press/v139/radford21a/radford21a.pdf

[4] Liu F., Ren R., Wu Y., et al. (2023). On the Compositional Generalization of Text-to-Image Diffusion Models. NeurIPS. https: //proceedings.neurips.cc/paper_files/paper/2024/file/b288470688e72f58c02031304ad6339f-Paper-Conference.pdf

[5] Whata A., Dibeco K., Madzima K. and Obagbuwa I. (2024) Uncertainty quantification in multi-class image classification using chest X-ray images of COVID-19 and pneumonia. Front. Artif. Intell. 7: 1410841. doi: 10.3389/frai.2024.1410841

[6] Liu F., Ren R., Wu Y., et al. (2024). Towards Understanding and Quantifying Uncertainty for Text-to-Image Generation. arXiv preprint. https: //dl.acm.org/doi/full/10.1145/3744238

[7] Zhang H., Li Y., & Wang Z. (2025). Trustworthy Evaluation of Generative AI Models. arXiv preprint. https: //arxiv.org/abs/2502.14296

[8] Gawlikowski J., et al. (2023). A Survey of Uncertainty in Deep Neural Networks. Artif Intell Rev 56 (Suppl 1), 1513–1589. https: //doi.org/10.1007/s10462-023-10562-9.

[9] Wei Y., et al. (2025). Personalized Image Generation with Deep Generative Models: A Decade Survey. https: //www.sciopen.com/article/10.26599/CVM.2025.9450495

[10] Nature News. (2024). When AI-generated images go wrong. https: //www.nature.com/articles/d41586-024-02420-7