Evaluating Generative Image Expansion for Long-Range Maritime Vision Tasks

Jaye Nias[1], Saurav K. Aryal[1], Joseph Sankah[1], Jeremy Blackstone[1], Armisha Roberts[1], Simone Smarr[1], Lucretia Williams[1], Gloria Washington[1]

[1]Human-Centered AI Institute, Howard University, Washington, DC, USA
jaye.nias@howard.edu, saurav.aryal@howard.edu, joseph.sankah@bison.howard.edu, jeremy.m.blackstone@howard.edu, armisha.roberts1@howard.edu, simone.smarr@howard.edu, lucretia.williams1@howard.edu, gloria.washington@howard.edu

Abstract

Synthetic image generation is increasingly used to augment visual datasets when real-world data is limited or difficult to capture. However, generative techniques do not simply extend existing images; they actively construct contextual assumptions about background continuity, spatial relationships, and scene structure. In decision-relevant settings, these assumptions can obscure uncertainty and introduce ambiguity that affects both model behavior and human interpretation. This paper examines the use of generative image expansion to induce distance-related perceptual stress in naval vessel imagery, motivated by the needs of maritime decision support under conditions aligned with Tactical Decision-Making Under Stress (TADMUS). Rather than evaluating downstream model performance, we focus on the interpretability and contextual integrity of augmented images as perceived by human annotators. We construct a dataset of perceptually degraded image variants using multiple generative platforms and assess output quality using a structured, context-focused annotation protocol.

Introduction

Tactical decision-making in maritime and naval environments frequently occurs under conditions of stress, uncertainty, and incomplete information. Operators must interpret sensor data, visual cues, and AI-supported assessments while facing time pressure and potentially high consequences. Within the framework of Tactical Decision-Making Under Stress (TADMUS) , the quality and interpretability of perceptual inputs are not peripheral concerns; they directly shape human judgment, trust calibration, and action selection (Morrison et al. 1996). Yet the visual data used to develop and assess AI-enabled perception systems often overrepresents idealized conditions, leaving degraded or ambiguous viewing scenarios insufficiently examined.

While synthetic image augmentation has advanced rapidly in recent years, its use for deliberately modeling perceptual stressors such as distance, occlusion, or diminished visual salience remains limited. Many applications of synthetic image augmentation prioritize visual diversity or realism, while the implications of perceptual degradation for downstream decision-making remain largely underexplored. Moreover, such practices are not yet systematically integrated into evaluation pipelines, particularly in decision-relevant or operational contexts. Generative techniques do not merely remove information; they actively construct new contextual assumptions about background continuity, spatial relationships, and scene structure. When these assumptions are not explicitly surfaced, AI systems trained or evaluated on such data may present outputs with unwarranted confidence, complicating tactical decision-making rather than supporting it. In this work, we frame synthetic augmentation and perceptual degradation as a form of perceptual stress induction relevant to TADMUS contexts. We present a controlled pipeline for inducing viewing-distance degradation in naval ship imagery and a context-rich annotation framework designed to characterize image integrity, contextual plausibility, and interpretability as decision-relevant signals.

Operational Motivation and Problem Context

Problem Statement

We seek to construct a training and evaluation dataset that supports a vision-to-text model for maritime vessel identification, with particular emphasis on distinguishing friendly from non-friendly naval vessels under distance-induced perceptual stress degradation. While publicly available imagery provides a practical starting point for prototyping, such data is uneven with respect to viewing distance and target salience. The problem addressed in this work is whether generative image expansion techniques can be used to produce distance-degraded variants that are sufficiently coherent, interpretable, and trustworthy to support downstream training and testing, and how the limitations of these techniques can be systematically characterized.

Context

This study emerged from an ongoing effort to develop a vision-to-text capability that supports an LLM-centered decision support workflow for maritime awareness. A core objective of this capability is to identify whether an observed naval vessel is friendly or non-friendly based on visual evidence and contextual cues. As an initial step, we constructed a seed image set by sampling frames from publicly available online videos. A researcher reviewed maritime footage

and extracted still images depicting naval vessels from multiple countries, producing a diverse but uneven dataset that reflects real-world variability in capture quality, viewpoint, and background conditions. A key limitation of the resulting dataset is that it overrepresents relatively clear views and mid-range observations, while providing limited coverage of long-range conditions where vessels occupy a small fraction of the frame and critical identifying features are less visible. Because distance and reduced target salience are central stressors for perception in maritime environments, we explored whether generative image expansion tools could be used to create distance-stressed variants of existing images while preserving vessel integrity and producing plausible maritime context.

This paper focuses on the strengths and limitations of several generative AI platforms for this specific augmentation task. Rather than treating all synthetic outputs as interchangeable, we evaluate how different tools behave under controlled distance-stress conditions, and we assess output quality through a human annotation protocol designed to capture subject integrity, dimensional fidelity, background plausibility, and artifact presence. The goal is to inform practical dataset construction decisions for training and testing vision-to-text models intended for decision-relevant maritime classification tasks.

Related Work

This work sits at the intersection of three bodies of scholarship: synthetic image augmentation, robustness and perceptual degradation in vision systems, and human-centered AI for decision support. While each of these areas has been studied extensively in isolation, they are rarely considered together in decision-relevant or operational contexts. In particular, existing work often separates questions of visual data generation from questions of interpretability, uncertainty, and human judgment under stress. This separation becomes especially salient in settings aligned with Tactical Decision-Making Under Stress (TADMUS) (Morrison et al. 1996), where perceptual inputs directly shape sensemaking, trust calibration, and action selection. The following sections situate this study within these literature and clarify how our focus on distance-induced perceptual stress degradation and contextual integrity extends existing work toward evaluation practices better aligned with operational decision-making needs.

Synthetic Image Augmentation and Visual Realism

Prior work on synthetic image augmentation has largely focused on increasing dataset diversity, improving class balance, or enhancing visual realism (Shorten and Khoshgoftaar 2019; Tremblay et al. 2018; Richter et al. 2016). Techniques such as image inpainting, scene expansion, and generative background synthesis are commonly evaluated using perceptual quality metrics or downstream model performance (Pathak et al. 2016; Rombach et al. 2022). In these settings, visual coherence and realism are often treated as proxies for data usefulness.

While such approaches have enabled scalable dataset construction, they typically do not make explicit the assumptions introduced during generative modification, nor do they characterize how these assumptions affect interpretability. Synthetic images are frequently treated as equivalent to real observations, despite embedding inferred or hallucinated context that may not be grounded in the original scene. As generative tools become more accessible, this lack of contextual accounting raises concerns about how synthetic data is incorporated into training and evaluation pipelines without sufficient attention to representational fidelity.

Perceptual Degradation and Robustness in Vision Systems

Research on robustness in computer vision has examined system behavior under perceptual degradations such as noise, blur, occlusion, and resolution loss (Hendrycks and Dietterich 2019; Geirhos et al. 2018; Dodge and Karam 2017). These studies have been instrumental in identifying failure modes and performance gaps, particularly under distributional shift. However, robustness evaluations are typically model-centric, emphasizing accuracy, calibration, or stability metrics rather than interpretability or downstream decision impact (Taori et al. 2020).

Less attention has been paid to how perceptual degradation is represented to human users or how uncertainty introduced at the perception stage propagates into decision-support workflows. In operational contexts, visually plausible outputs may obscure meaningful uncertainty, particularly when generative techniques introduce coherent but ungrounded context. This gap motivates evaluation approaches that consider not only whether models remain accurate under degradation, but whether degraded or augmented inputs remain interpretable and trustworthy to human decision-makers.

Human-Centered AI and Decision-Support Contexts

Human-centered AI research emphasizes the importance of transparency, interpretability, and appropriate trust calibration in AI-assisted decision-making, particularly in high-stakes and time-pressured environments (Amershi et al. 2019; Hoffman, Klein, and Mueller 2018; Suresh and Guttag 2021). Foundational work in situation awareness and naturalistic decision-making highlights how perceptual cues and uncertainty representation shape human judgment under stress (Endsley 2017; Klein 2017).

Within the Tactical Decision-Making Under Stress (TADMUS) framework, perceptual inputs are not neutral; they actively influence sensemaking, confidence, and action selection under conditions of uncertainty and time pressure. Despite this, relatively little work has examined how synthetic visual data and generative preprocessing affect contextual integrity and uncertainty representation in decision-support systems. When generative techniques are used to augment or preprocess visual inputs, the resulting context is rarely annotated or evaluated in ways that reflect decision relevance. This gap motivates approaches that treat perceptual stress, contextual fidelity, and human interpretability as first-class evaluation concerns.

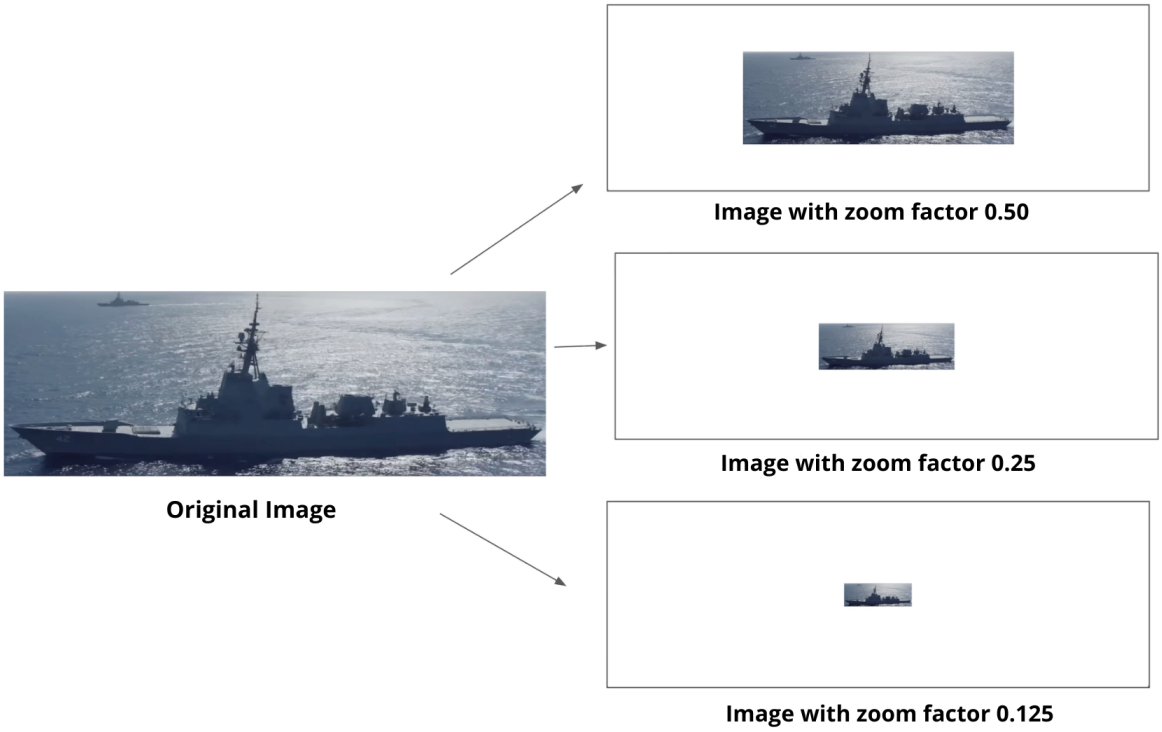

Figure 1: Original image vs. Images zoomed out and padded

Methods

Our methodological approach is designed to evaluate whether generative image expansion techniques can be used to produce distance-stressed variants of naval vessel imagery that remain interpretable and decision-relevant. Rather than optimizing for visual realism alone, we characterize how different generative platforms behave under controlled perceptual stress and how the resulting images preserve or distort contextual cues relevant to vessel identification. The methods described below detail the construction of the seed dataset, the procedure for inducing distance-based perceptual stress, the generative platforms evaluated, and the annotation framework used to assess output fidelity and interpretability.

Dataset Construction

The seed dataset was constructed from publicly available online videos depicting naval vessels from multiple countries. A researcher manually reviewed maritime footage and extracted still frames containing identifiable naval ships, resulting in an initial collection of approximately 100 images. This dataset reflects real-world variability in capture conditions, including differences in resolution, camera angle, background clutter, and environmental context. While the dataset provides broad coverage of vessel types and viewing conditions, it is not uniformly distributed with respect to viewing distance. In particular, long-range observations in which vessels occupy a small portion of the image are underrepresented. This imbalance motivated the use of synthetic augmentation to induce distance-related perceptual stress in a controlled manner.

Distance-Stress Induction via Generative Image Expansion

To simulate distance-induced perceptual stress, we generated distance-stressed variants by reducing the apparent size of the target vessel within the image while preserving its spatial location. Operationally, this was implemented by resizing the original image and embedding it within an expanded canvas, thereby increasing the proportion of background relative to the vessel.

Three levels of perceptual degradation were applied, corresponding to progressively reduced target salience. Figure 1 illustrates the distance-stress induction process, showing the original vessel image alongside progressively reduced target-salience variants prior to generative background expansion. Following this resizing step, generative image expansion was used to synthesize background content surrounding the resized image. This approach allowed us to assess how well generative techniques could maintain subject integrity and plausible maritime context as perceptual stress increased.

Generative Platforms Evaluated

We evaluated multiple generative AI platforms commonly used for image generation or modification, including web-based tools, API-driven services, and application-based generative editors. Table 1 summarizes the platforms evaluated

Platform	Access	Batch	Stability
ChatGPT+	Subscription	No	Unstable
Midjourney	Subscription	No	Unstable
Vertex AI Studio	Usage-based	Yes	Unstable
Meta LLaMA	Self-hosted	No	Unstable
Photoshop Gen Fill	Subscription	Yes	Very Stable

Table 1: Operational characteristics of generative image expansion platforms evaluated.

and key operational characteristics relevant to dataset construction. These platforms were selected to reflect a range of cost structures, interaction paradigms, and degrees of user control. For each platform, we assessed its ability to perform generative image expansion following distance-stress induction, as well as the stability and consistency of the resulting outputs. Particular attention was paid to background coherence, artifact introduction, and the preservation of vessel geometry across repeated generations.

Annotation Protocol

To evaluate the quality and interpretability of perceptually degregraded images, we employed a human annotation protocol designed to capture contextual fidelity rather than aesthetic preference. Each augmented image was assigned a categorical quality label reflecting its suitability for decision-relevant use in a maritime identification context. The protocol is designed to surface ambiguity and failure modes that are meaningful for downstream decision-making, rather than to rank images by visual appeal.

Annotations focused on five dimensions: subject integrity, subject dimensional fidelity, background quality, background consistency, and artifact presence.

- *Subject integrity* refers to whether the vessel remains a coherent and identifiable object rather than being distorted or merged with the background.
- *Subject dimensional fidelity* captures whether the relative proportions and spatial extent of the vessel are preserved under augmentation.
- *Background quality* reflects the visual plausibility of the generated environment.
- *Background consistency* assesses whether the background aligns contextually with the vessel and scene.
- *Artifact presence* captures the introduction of unnatural or implausible visual elements resulting from generative intervention.

These dimensions were selected to reflect whether the augmented image preserved critical visual cues needed for vessel identification without introducing misleading or implausible contextual elements. To support consistent application of these dimensions, annotators used a three-level ordinal scale reflecting overall image suitability for decision-relevant use. Each level corresponds to characteristic patterns of subject distortion, contextual inconsistency, and artifact presence, as summarized in Table 2.

Quality Level	Characteristic Attributes
Poor (Red)	At least one major failure mode is present, including severe subject distortion or merging with the background, substantial loss of subject dimensions, inconsistent or implausible background context, or prominent generative artifacts that introduce misleading visual elements.
Moderate (Yellow)	Partial degradation is present, such as mild subject distortion or reduced dimensional fidelity, moderate background inconsistency, or minor but noticeable generative artifacts that may introduce ambiguity without fully obscuring vessel identity.
High (Green)	Subject integrity and dimensions are preserved, background context is visually plausible and consistent with the vessel, and no noticeable generative artifacts are present. Images in this category support reliable interpretation under distance-induced perceptual stress.

Table 2: Annotation scale used to assess the quality and interpretability of distance-stressed images.

Figure 2 provides representative examples for each annotation category, illustrating how different degrees of distance-induced perceptual stress and generative intervention manifest in practice. The figure highlights characteristic failure modes observed in lower-quality images, such as subject deformation, background inconsistency, and unnatural artifacts, as well as examples where subject integrity and contextual plausibility are preserved under moderate stress. These examples served as shared reference points during annotation to promote consistency in how quality criteria were interpreted.

Inter-Rater Reliability Analysis

Two independent annotators evaluated all augmented images. To assess consistency and identify sources of ambiguity, we computed multiple inter-rater reliability metrics, including Cohen's kappa, Spearman's rank correlation, and Krippendorff's alpha. These measures provide complementary perspectives on agreement across nominal and ordinal interpretations of the annotation scale. Rather than treating disagreement as noise, we interpret reduced agreement under higher perceptual degradation as an indicator of representational instability introduced by generative augmentation. These analyses inform both the limits of the augmentation approach and the reliability of the resulting dataset for downstream use.

Results

This section reports the outcomes of the perceptual degradation augmentation and annotation process. Rather than evaluating model accuracy or downstream classification performance, the results characterize patterns in image quality, annotation agreement, and representational stability across

Tag Condition	Sample Image
Image has at least one of these attributes: **Subject Integrity**: severely altered, merged with the background or object **Subject Dimensions**: severely lost **Background Quality**: High **Background consistency**: inconsistent **Background artifacts**: major unnatural modifications (objects that don't belong)	 Gen-fill image w/ resize factor 0.25
Image has at least one of these attributes: **Subject Integrity**: partially altered **Subject Dimensions**: partially lost **Background quality**: moderate **Background Consistency**: partially consistent **Background artifacts**: minor noticeable imperfections	 Gen-fill image w/ manual task w/ resize factor 0.125
Image has all these attributes: **Subject Integrity**: Intact **Subject Dimensions**: dimensions preserved **Background Quality**: high **Background consistency**: consistent with the subject **Background artifacts**: No artifacts/natural	 Gen-fill image w/ input prompt and resize factor 0.50

Figure 2: Representative examples for each annotation category (poor, moderate, high), illustrating subject integrity, background consistency, and artifact presence under distance-induced perceptual stress.

distance-stress conditions and generative configurations. We employ inter-rater agreement and dimension specific ratings as proxies for interpretability and representational reliability. In this context, agreement reflects the extend to which an image affords consistent semantic interpretation under degraded conditions. These measures are commonly used when ground truth is ambigous but interpretive consistency is operationally meaningful.

Dataset Scale and Augmentation Coverage

Starting from an initial seed set of approximately 100 images, the augmentation pipeline produced a total of 882 distance-stressed variants. These images span three levels of distance-induced perceptual stress and multiple generative configurations, yielding a dataset that systematically varies target salience while holding the underlying vessel identity constant. All augmented images were annotated using the three-level ordinal quality scale described in Table 2.

Image Quality Distribution Across Perceptual Degradation

Annotation outcomes show that increasing perceptual degradation is associated with greater difficulty in interpreting augmented images. Agreement between annotators is higher for images generated under moderate distance-stress conditions and drops substantially for the most extreme stress condition. For example, measures of agreement such as Spearman's ρ decrease from values near 0.50 under moderate stress to approximately 0.21 at the highest stress level, while overall agreement measures fall below 0.20 for the most aggressive distance condition.

These shifts indicate that as vessels occupy a smaller portion of the image and generative expansion plays a larger role in constructing the surrounding context, annotators become less consistent in how they assess image quality. This reduced agreement aligns with observed degradation in subject integrity and background coherence, suggesting that higher perceptual degradation introduces ambiguity that affects both visual fidelity and human interpretation.

Grouping Criterion	N	nominal α	ordinal α	κ (unweighted)	κ (linear)	κ (quadratic)	Spearman's ρ
Overall	486	0.40	0.50	0.40	0.45	0.50	0.50
Gen fill	292	0.39	0.47	0.39	0.44	0.48	0.48
Gen fill + Prompt	194	0.38	0.50	0.38	0.45	0.50	0.51
Resize = 0.5	194	0.42	0.51	0.42	0.46	0.50	0.51
Resize = 0.25	194	0.37	0.52	0.37	0.46	0.53	0.53
Resize = 0.125	98	0.17	0.16	0.19	0.19	0.19	0.21

Table 3: Inter-rater reliability measures with 2 independent raters. Spearman's p-value ≤ 0.05.

Inter-Rater Reliability Under Increasing Perceptual Stress

Inter-rater reliability outcomes are summarized in Table 3 and provide insight into how consistently distance-stressed images could be interpreted using the shared annotation protocol. Across the full dataset ($N = 486$), overall agreement falls in the moderate range (e.g., ordinal $\alpha = 0.50$, Spearman's $\rho = 0.50$) with substantial variation across perceptual degradation conditions.

Agreement was highest for images generated under moderate perceptual degradation and declined as perceptual stress increased. This pattern is visible across all reported measures in Table 3. For example, values for Spearman's ρ and Krippendorff's α are near 0.50 for moderate stress conditions, but drop to approximately 0.20 or lower for the most extreme degradation condition (e.g., ordinal $\alpha = 0.16$, Spearman's $\rho = 0.21$ at Resize = 0.125), where vessels occupy only a small portion of the image. Similar declines are observed for Cohen's κ, indicating that reduced agreement is not specific to a single metric.

These patterns suggest that as perceptual degradation increases, augmented images become more difficult to interpret in a consistent way, even when annotators share a common scale and reference examples. Rather than reflecting annotation error alone, the decline in agreement points to increased ambiguity in the visual representations themselves. In this sense, inter-rater reliability serves as a practical indicator of representational stability, highlighting where generative image expansion begins to introduce uncertainty that may limit the usefulness of augmented images for decision-relevant tasks. While these agreement measures are not intended as formal tests of statistical significance, they offer a useful lens for understanding how perceptual stress affects the interpretability of generative outputs.

Beyond overall trends, several additional patterns emerge from the inter-rater agreement data. Agreement measures that account for ordinal relationships consistently exceed nominal agreement, suggesting that annotators often differed by degree rather than by fundamental judgment. Prompt-based generative expansion did not yield a clear improvement in agreement, indicating that simple prompt guidance may not mitigate ambiguity introduced by perceptual degradation. Finally, the sharp decline in agreement at the highest stress level suggests a threshold beyond which generative image expansion produces representations that are no longer reliably interpretable.

Discussion

The results highlight both the promise and the limits of generative image expansion for constructing distance-stressed visual datasets in decision-relevant maritime contexts. Taken together, the findings suggest that the behavior of generative augmentation under perceptual stress is neither uniformly beneficial nor uniformly harmful. Instead, its utility depends on the degree of stress applied and on whether distortions introduced during generation are made visible and interpretable within the dataset. The discussion below surfaces several grounded observations that emerge from this analysis and considers their implications for decision-relevant AI systems.

Generative image expansion exhibits a narrow window of reliability under perceptual stress. Inter-rater agreement remains relatively stable under low to moderate perceptual degradation but drops sharply at the highest stress level, suggesting a threshold beyond which generative expansion produces representations that are no longer consistently interpretable. Under moderate perceptual stress, generative techniques were often able to preserve subject integrity and plausible background context, producing images that annotators judged as suitable for interpretive use. However, as perceptual degradation increased, generative assumptions about background continuity and spatial structure became more pronounced, leading to representational distortions that undermined interpretability. This pattern suggests that generative expansion should not be treated as a drop-in solution across all stress conditions, particularly when extreme reductions in target salience are involved.

Prompt-based generative guidance does not reliably mitigate ambiguity under perceptual degradation. While prompting was used to encourage plausible maritime backgrounds, agreement patterns indicate that simple prompt guidance does not substantially improve interpretability once perceptual degradation increases. This is reflected in the agreement metrics, which are nearly identical for generative fill with and without prompting (e.g., ordinal $\alpha = 0.47$ vs. 0.50, Spearman's $\rho = 0.48$ vs. 0.51; Table 3). This suggests that perceptual degradation, rather than prompt specificity, is the dominant factor shaping how generative outputs are interpreted.

Synthetic visual data functions as an intervention, not a neutral transformation. From a tactical decision-making perspective, these findings underscore the importance of treating synthetic visual data as an active intervention in the perceptual pipeline rather than as a simple extension of existing data. Generative expansion does not merely reduce or

obscure information; it introduces inferred context that may mask uncertainty or fabricate coherence where none exists. In settings aligned with Tactical Decision-Making Under Stress (TADMUS), such distortions pose a risk to appropriate trust calibration, particularly when AI-generated visual inputs are consumed alongside model-generated textual assessments.

Human disagreement under perceptual degradation surfaces meaningful signals of ambiguity. Notably, agreement measures that account for ordinal relationships remain higher than nominal agreement across most conditions. For example, across most conditions ordinal α exceeds nominal α by approximately 0.08–0.15 (Table 3), indicating differences in degree rather than categorical disagreement. The sharp decline in agreement at high degradation levels, however, reflects a breakdown in representational stability rather than simple boundary disagreement.

Disciplined use of synthetic imagery requires explicit indicators of fidelity and limitation. Importantly, this work does not argue against the use of generative image expansion in dataset construction. Instead, it motivates a more disciplined approach in which distance-degradated variants are accompanied by explicit indicators of fidelity, ambiguity, and limitation. Context-rich annotation and reliability analysis provide practical mechanisms for bounding the applicability of synthetic imagery and for preventing augmented data from propagating unwarranted confidence into downstream vision-to-text and decision-support workflows.

Limitations

This study is limited to a single visual domain and a small, intentionally constrained seed dataset derived from publicly available imagery. Inter-rater agreement is moderate overall, but is used analytically as a signal of interpretability under perceptual degradation rather than as a measure of annotator performance. Constraints on image fidelity, distance, and target size reflect controlled design parameters rather than data scarcity. Finally, because the analysis focuses on generative augmentation behavior rather than model comparison, findings may vary across generative model architectures and versions.

Conclusion and Future Work

This work contributes to the study of AI in practice by examining the strengths and limitations of generative image expansion as a dataset construction strategy in decision-relevant settings. Rather than asserting improvements in model performance, the findings highlight how synthetic visual interventions shape interpretability, ambiguity, and trust signals that are critical for real-world use. In maritime and defense-oriented contexts, where visual inputs are consumed under uncertainty and time pressure, understanding these effects is essential. By framing distance-degraded image generation as a form of perceptual stress induction, this research surfaces risks associated with unexamined generative assumptions, including representational distortion and unwarranted confidence. The results underscore the need for context-rich annotation and reliability analysis when integrating synthetic visual data into AI-enabled workflows. In doing so, this work advances a more disciplined and human-centered approach to synthetic data use, supporting responsible innovation without overstating capability or performance gains.

Acknowledgments

This material is based upon work supported by the Office of Naval Research under Grant No. N00014-22-1-2714. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Office of Naval Research, the Department of Defense or the Department of War.

References

Amershi, S.; Weld, D.; Vorvoreanu, M.; Fourney, A.; Nushi, B.; Collisson, P.; Suh, J.; Iqbal, S.; Bennett, P. N.; Inkpen, K.; et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*, 1–13.

Dodge, S.; and Karam, L. 2017. A study and comparison of human and deep learning recognition performance under visual distortions. In *2017 26th international conference on computer communication and networks (ICCCN)*, 1–7. IEEE.

Endsley, M. R. 2017. Toward a theory of situation awareness in dynamic systems. In *Situational awareness*, 9–42. Routledge.

Geirhos, R.; Temme, C. R.; Rauber, J.; Schütt, H. H.; Bethge, M.; and Wichmann, F. A. 2018. Generalisation in humans and deep neural networks. *Advances in neural information processing systems*, 31.

Hendrycks, D.; and Dietterich, T. 2019. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*.

Hoffman, R. R.; Klein, G.; and Mueller, S. T. 2018. Explaining explanation for "explainable AI". In *Proceedings of the human factors and ergonomics society annual meeting*, volume 62, 197–201. SAGE Publications Sage CA: Los Angeles, CA.

Klein, G. A. 2017. *Sources of power: How people make decisions*. MIT press.

Morrison, J. G.; Kelly, R. T.; Moore, R. A.; and Hutchins, S. G. 1996. Tactical decision making under stress (TADMUS) decision support system.

Pathak, D.; Krahenbuhl, P.; Donahue, J.; Darrell, T.; and Efros, A. A. 2016. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2536–2544.

Richter, S. R.; Vineet, V.; Roth, S.; and Koltun, V. 2016. Playing for data: Ground truth from computer games. In *European conference on computer vision*, 102–118. Springer.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Shorten, C.; and Khoshgoftaar, T. M. 2019. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1): 1–48.

Suresh, H.; and Guttag, J. 2021. A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–9.

Taori, R.; Dave, A.; Shankar, V.; Carlini, N.; Recht, B.; and Schmidt, L. 2020. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33: 18583–18599.

Tremblay, J.; Prakash, A.; Acuna, D.; Brophy, M.; Jampani, V.; Anil, C.; To, T.; Cameracci, E.; Boochoon, S.; and Birchfield, S. 2018. Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 969–977.