Uncertainty-of-Information-Driven GAN (UoI GAN): Quantifying and Communicating Uncertainty to Decision-Makers

Anjon Basak,[2]Rajendran Swamidurai,[3]Adrienne Raglin[1]

[1]DEVCOM Army Research Laboratory
[2]Oak Ridge Associated Universities
[3]Alabama State University

Abstract

Generative Adversarial Networks (GANs) are cutting-edge machine learning algorithms that can generate realistic data such as images and time series. Their application has been explored in autonomous systems, fraud detection, and medical diagnostics. However, the "black box" nature and intrinsic instability of these models pose significant and frequently hidden risks. Decision-makers struggle to interpret GAN outputs because, unlike predictive models, GANs do not explicitly measure confidence. Relying solely on aggregate performance metrics such as the Fréchet Inception Distance can obscure important failure modes. GAN systems must therefore be designed to explicitly expose and communicate their uncertainty. In this work, we propose a hybrid model that combines explicit uncertainty representation with quantitative uncertainty-of-information measurement. The proposed framework redesigns the GAN architecture to intrinsically represent uncertainty by examining output variance. It further provides a principled mechanism for translating these complex metrics into interpretable signals that expose model behavior and limitations.

Introduction

Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) have significantly influenced the domain of artificial intelligence, exhibiting a strong ability to generate highly realistic data (Amazon Web Services 2025). Synthetic data produced by GANs is widely used across multiple fields. In medicine, for example, GAN-generated samples have been explored as a mechanism for protecting patient privacy while enabling downstream analysis (Ramachandranpillai et al. 2024).

GANs have been applied extensively across domains that are characterized by limited data availability, high acquisition costs, or sensitivity to privacy concerns (Amazon Web Services 2025). Application areas include photorealistic image generation, image-to-image translation, video and 3D content generation, medical imaging, drug discovery and molecular design, art and design, fashion and e-commerce, and video game development.

GANs can produce new images from text prompts or modify existing images, enabling realistic and immersive visual experiences for digital entertainment (Amazon Web Services 2025). In these cases, GANs generate new data points intended to capture the underlying patterns of the original dataset. This capability is commonly leveraged to improve data availability for subsequent machine learning tasks (Amazon Web Services 2025).

The GAN architecture consists of two neural networks—a generator and a discriminator—that compete in a zero-sum game (Goodfellow et al. 2014). This adversarial mechanism allows GANs to implicitly learn complex, high-dimensional data distributions across images, audio, and other modalities, supporting applications such as realistic image synthesis, synthetic training data generation, and data completion (Amazon Web Services 2025).

Despite their ability to generate highly realistic outputs, GANs present substantial challenges when deployed in high-stakes domains (Vuletić, Cucuringu, and Prenzel 2023). Conventional deep learning systems, including GANs, typically produce single-point estimates without explicitly representing model confidence or potential inaccuracy. In critical domains such as medicine, finance, and autonomous systems, reliance on single-point generative outputs limits the ability to assess failure modes and model reliability (Vuletić, Cucuringu, and Prenzel 2023).

The concept of Uncertainty of Information (UoI) provides a framework for characterizing model confidence and the range of possible outcomes. Transforming qualitative statements such as "this model may be erroneous" into precise, quantifiable representations of uncertainty is essential in contexts where trustworthiness is critical (IBM 2025).

In this research, we investigate three central questions regarding uncertainty in GAN-based systems:

1. How does uncertainty inherent in GAN architecture and training manifest in generative behavior?
2. How can uncertainty be measured using state-of-the-art technical methods?
3. How can uncertainty be explicitly represented within GAN models to expose model behavior and limitations?

We address these questions by outlining the foundational GAN architecture, analyzing its intrinsic instability through the lens of Uncertainty of Information, and proposing an uncertainty-aware modification to the discriminator that enables uncertainty to be surfaced as a first-class signal within

the generative process.

Background

Fundamentals of Generative Adversarial Networks (GANs)

A Generative Adversarial Network (GAN) is a deep learning framework composed of two adversarial neural networks: a generator (G) and a discriminator (D). These networks compete in a zero-sum game in which each model is trained to outperform the other. The generator receives a random noise vector as input and learns to produce synthetic data samples that approximate the distribution of the real training data. Its objective is to deceive the discriminator into classifying generated samples as authentic. The discriminator, in turn, is trained to distinguish between genuine samples from the training dataset and synthetic samples produced by the generator. It seeks to minimize classification error and improve its ability to correctly identify real versus generated data (Amazon Web Services 2025).

The adversarial training process proceeds iteratively, with both networks updating their parameters in alternating fashion until a form of equilibrium is reached. At convergence, the generator ideally produces outputs that are sufficiently realistic such that the discriminator can no longer reliably distinguish them from true data samples (Amazon Web Services 2025).

A common challenge in GAN training is *mode collapse*. Mode collapse occurs when the generator learns to reproduce only a limited subset of the training distribution, producing a narrow range of outputs rather than capturing the full diversity of the target data distribution (Saatchi and Wilson 2017). This limitation motivates our proposed framework, as certain probabilistic GAN formulations address this issue by modeling a distribution over generators rather than relying on a single point estimate (MIT CSAIL 2018).

The Unique Challenge of GANs

Generative Adversarial Networks (GANs) encounter challenges that are not present in traditional predictive models, primarily due to their adversarial training paradigm. The two core components of a GAN—the generator network (G) and the discriminator network (D)—compete in a setting where one network's gain corresponds to the other's loss. The generator learns to produce synthetic data that is indistinguishable from real data, while the discriminator learns to differentiate between authentic and generated samples. This dynamic interaction aims to reach an equilibrium; however, in practice, it frequently exhibits instability. One prevalent and subtle failure mode associated with this instability is *mode collapse* (Chan, Molina, and Metzler 2024).

Mode collapse occurs when the generator fails to capture the full diversity of the training data distribution, instead producing a limited set of repeating or nearly identical outputs (Spot Intelligence 2023). For example, a GAN trained on a heterogeneous image dataset may generate only variations of a single object—such as one specific breed of dog—rather than producing the intended diversity of distinct objects. This phenomenon undermines the fundamental objective of generative modeling, which is to produce a rich and diverse distribution of samples (Spot Intelligence 2023).

The root cause of this issue lies in an imbalance during adversarial training. The generator may discover a narrow set of outputs that consistently deceive the discriminator, while the discriminator fails to learn the complete structure of the underlying data distribution. As a result, the adversarial process converges to a suboptimal equilibrium that masks the loss of diversity in the generated outputs (Spot Intelligence 2023).

Foundational Concepts of Uncertainty of Information (UoI)

Uncertainty of Information (UoI) is a broad foundational concept referring to the general condition of imperfect or unknown information. The concept extends beyond purely probabilistic interpretations to encompass more nuanced forms of imperfect knowledge.

Several distinct forms of uncertainty fall under this umbrella. *Vagueness* refers to the inability to clearly distinguish between conceptual categories, such as "person of average height" versus "tall person". *Ambiguity* arises when possible outcomes admit multiple interpretations, as in the case of the word "bank," which may refer to a financial institution or a riverbank.

Within the broader UoI framework, uncertainty may manifest across different modalities of data, devices, and systems originating from diverse sources. *Incompleteness* occurs when required data or information is missing from a dataset. *Corruption* refers to the presence of errors within data. *Inconsistency* describes discrepancies relative to previously known states or established data. *Inappropriateness* arises when data or system states are outdated or unsuitable for a given process.

Uncertainty is inherently linked to risk, which plays a central role in decision-making tasks. A key distinction, articulated by Frank Knight, separates *risk* (measurable probability) from *true uncertainty* (immeasurable lack of knowledge), including the presence of "unknown unknowns". The UoI framework therefore provides flexibility to incorporate both quantifiable risk and non-quantifiable forms of uncertainty.

An additional form particularly relevant for future exploration is *representational uncertainty*, which describes the conceptual gap between a real-world construct (e.g., "video quality") and its operational measurement (e.g., "Like Rate"). Within machine learning contexts, UoI connects to both *aleatoric uncertainty*, arising from inherent variability in data, and *epistemic uncertainty*, which stems from incomplete knowledge about the model or environment.

The Causal Relationship Between GAN Failures and Uncertainty

The phenomenon of mode collapse is not merely a technical flaw; it can be interpreted as a manifestation of a specific form of uncertainty. In this scenario, the generator fails to

learn the full data distribution and remains ignorant of entire modes or clusters that it is intended to model (Chan, Molina, and Metzler 2024). This reflects a form of model uncertainty that may be mitigated by modifying the training process to encourage broader coverage of the data distribution. Conversely, the use of a random noise vector z, which the generator maps to a sample, serves as a fundamental mechanism for introducing variability into the generative process and reflects uncertainty inherent to the underlying stochastic computation.

Distinguishing between these types of uncertainty provides a structured framework for analyzing GAN behavior. When a model exhibits elevated uncertainty, an essential step is identifying its source. Uncertainty arising from incomplete or insufficient information—such as underrepresented regions of the data distribution—differs from uncertainty caused by corrupt or noisy information, which may be inherently irreducible. Framing these phenomena within the Uncertainty of Information (UoI) perspective enables ambiguous behaviors, such as occasional instability or failure, to be described more precisely as elevated uncertainty within specific regions of the learned distribution. This reframing facilitates systematic analysis of failure modes and reveals where the generative process lacks sufficient constraint. It further motivates architectural and training modifications designed to expose or model such uncertainty, for example through improved characterization of latent space structure (Bayat 2023).

Prior research has investigated integrating GANs with uncertainty quantification (UQ) techniques (Ma 2024). Some studies have leveraged GANs to learn prior distributions over complex parameters within Bayesian inference frameworks (MIT CSAIL 2018), addressing challenges in prior specification by operating within a lower-dimensional latent space. Other work (Mirza and Osindero 2014) has demonstrated that Conditional GANs (cGANs) can function as predictive probabilistic models by reversing inputs and outputs to learn a distribution over labels, where the variance of that distribution can serve as an uncertainty signal. Collectively, these efforts reflect increasing recognition that explicitly modeling uncertainty is essential for analyzing and understanding generative models. These challenges directly motivate the perspective adopted in this work.

The UoI-Driven GAN Framework: Architecture

The proposed framework introduces a modification to the traditional GAN architecture by incorporating a novel *Uncertainty of Information-Aware Discriminator (UoI-AD)* at its core, forming what we denote as the *UoI-GAN*. The key architectural distinction is that the discriminator produces two distinct outputs, enabling a more informative and structured feedback loop for the generator. This design is inspired by prior work on Bayesian GANs and uncertainty-aware discriminators in related domains (Vuletić, Cucuringu, and Prenzel 2023).

Core Rationale

The central rationale of the UoI-GAN framework is to decompose GAN failures into measurable uncertainty components and to reformulate the adversarial objective accordingly. Rather than focusing solely on generating realistic samples that deceive the discriminator, the UoI-GAN seeks to produce outputs that are both realistic and consistently reliable, exhibiting low uncertainty across subgroups of the data distribution.

In the first phase of the framework, the discriminator functions not only as a classifier but also as a monitor of confidence and reliability. By decomposing total uncertainty into interpretable components, the framework aims to provide additional insight to support decision-making and downstream tasks. A component of the Uncertainty of Information (UoI) is associated with this confidence and reliability signal. The following comparative analysis positions the UoI-GAN within the broader landscape of generative models.

The Uncertainty of Information-Aware Discriminator (UoI-AD)

The principal innovation of the UoI-GAN framework is the introduction of the *Uncertainty of Information-Aware Discriminator (UoI-AD)* (see Figure 1). Unlike a conventional discriminator that outputs a single real/fake probability, the UoI-AD provides a dual output for each input sample:

- A standard adversarial score, $D_{\text{real/fake}}(x)$
- An uncertainty score or map, $D_{\text{uncertainty}}(x)$

The uncertainty output quantifies the discriminator's confidence in its classification. This representation may take the form of a scalar per-sample uncertainty score or a structured per-pixel confidence map, similar to approaches explored in super-resolution GAN frameworks (Vuletić, Cucuringu, and Prenzel 2023). This dual-output structure enables a richer feedback signal to the generator.

Conceptually, the UoI-AD serves as an adaptive internal monitor within the network. Its design is based on the premise that uncertainty can function as a proxy for quality. By identifying and penalizing regions of ambiguity or insufficient support in the learned distribution, the UoI-AD acts as an internal diagnostic signal within the adversarial process.

Dual-Objective Loss Functions for Stable and Reliable Training

Training of the UoI-GAN is guided by composite loss functions that extend the standard adversarial objective by incorporating uncertainty signals. The intent is not to guarantee stability but to explicitly integrate uncertainty into adversarial learning as an observable and responsive component.

The Discriminator's Objective: Representing Uncertainty The discriminator is trained using a composite objective:

$$L_D = L_{\text{adv}} + \lambda L_{\text{uncert}} \tag{1}$$

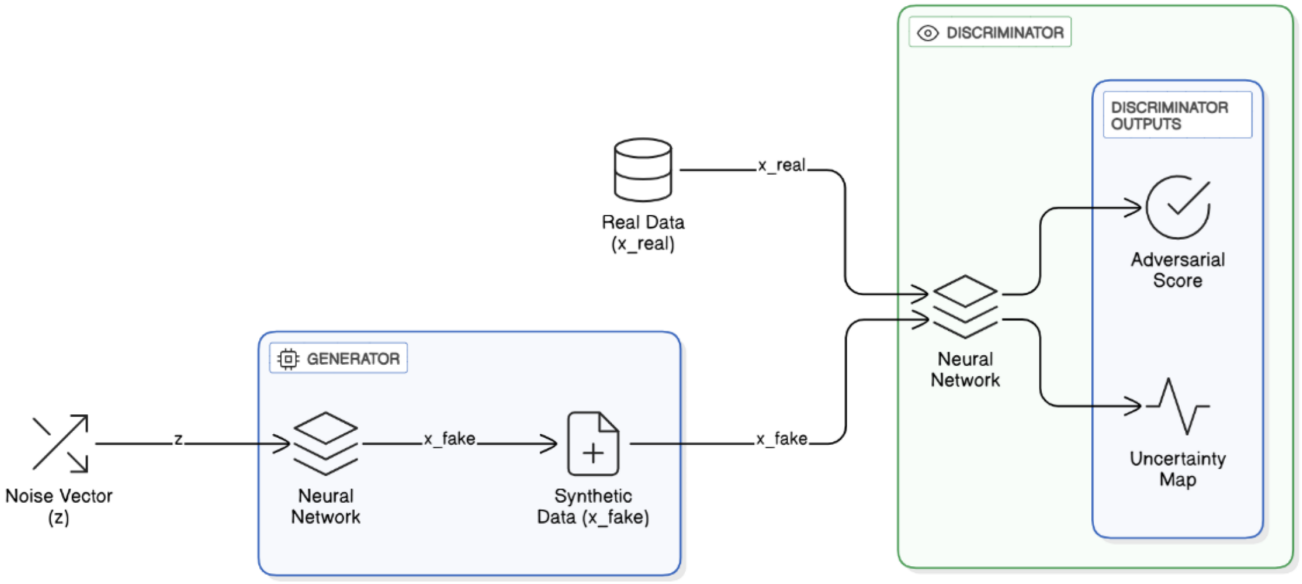

Figure 1: UoI-GAN Architecture Diagram Description.

Here, L_{adv} represents the conventional adversarial loss (e.g., binary cross-entropy), which trains the UoI-AD to distinguish between real and generated samples. The term L_{uncert} represents an uncertainty-related loss encouraging lower uncertainty for well-supported samples and higher uncertainty for ambiguous or weakly supported samples. The parameter λ controls the influence of the uncertainty component.

Thus, the discriminator is trained not only to classify but also to represent its confidence in that classification.

The Generator's Objective: Responding to Uncertainty Signals The generator is trained with a corresponding composite objective:

$$L_G = L'_{\text{adv}} + \beta L'_{\text{uncert}} \tag{2}$$

Here, L'_{adv} is the conventional adversarial objective encouraging the generator to produce samples classified as real. The term L'_{uncert} is derived from the uncertainty signal predicted by the UoI-AD for generated samples:

$$L'_{\text{uncert}} \propto D_{\text{uncertainty}}(x_{\text{fake}})$$

Minimizing this term encourages the generator to produce outputs associated with lower discriminator uncertainty, aligning generation with regions of the data distribution that are consistently supported. The parameter β balances adversarial and uncertainty-driven objectives.

Taken together, these objectives illustrate how uncertainty of information may be embedded directly into adversarial training dynamics. Rather than optimizing solely for realism, the generator–discriminator interaction explicitly incorporates uncertainty as a first-class signal within the learning process.

Conclusion

Generative Adversarial Networks represent a powerful class of models capable of producing highly realistic synthetic data, but their widespread adoption continues to be constrained by the opaque nature of their generative process. In particular, uncertainty is typically implicit, unmeasured, and difficult to reason about within standard GAN architectures. This paper argues that relying solely on black-box adversarial optimization is insufficient for understanding when and how generative models fail, especially in complex or high-stakes settings. Rather than proposing a fully validated system, this work presents a conceptual reframing of GAN behavior through the lens of Uncertainty of Information. By explicitly representing uncertainty within the discriminator and incorporating it into the adversarial learning process, the proposed framework illustrates how uncertainty can be surfaced as a first-class signal during generation. The intent is not to claim improvements in generative quality, stability, or coverage, but to demonstrate how uncertainty-aware mechanisms can make model behavior and limitations more explicit. This perspective also highlights several limitations. The uncertainty signals described in this framework are learned representations and may not correspond directly to formal notions of epistemic or aleatoric uncertainty. Furthermore, incorporating uncertainty into adversarial objectives does not eliminate known GAN pathologies, nor does it guarantee convergence or robustness. The proposed loss formulations and architectural modifications should therefore be viewed as illustrative rather than prescriptive. Over-

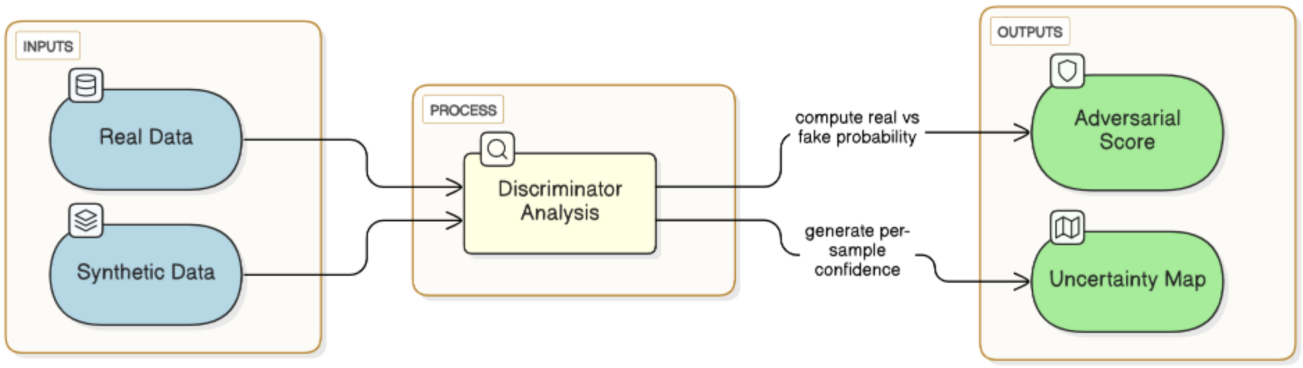

Figure 2: Uncertainty-of-Information-Aware Discriminator (UoI-AD)

all, this work positions uncertainty not as a post hoc diagnostic, but as an integral component of generative model design. By framing uncertainty as something to be represented and examined, rather than implicitly absorbed into optimization, the paper aims to encourage further research into uncertainty-aware generative models and more transparent interpretations of their behavior.

Acknowledgments

Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number **47QFCA21F0042**. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

Amazon Web Services. 2025. What Is a GAN? Generative Adversarial Networks Explained. Accessed August 29, 2025.

Bayat, R. 2023. A Study on Sample Diversity in Generative Models: GANs vs. Diffusion Models. In *International Conference on Learning Representations*.

Chan, M. A.; Molina, M. J.; and Metzler, C. A. 2024. Estimating Epistemic and Aleatoric Uncertainty with a Single Model. In *Advances in Neural Information Processing Systems*. 38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Networks. arXiv:1406.2661.

IBM. 2025. Uncertainty Quantification. Accessed August 22, 2025.

Ma, C. 2024. Uncertainty-Aware GAN for Single Image Super Resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Mirza, M.; and Osindero, S. 2014. Conditional Generative Adversarial Nets. arXiv:1411.1784.

MIT CSAIL. 2018. Bayesian Modelling and Monte Carlo Inference for GAN. Accessed August 22, 2025.

Ramachandranpillai, R.; Sikder, M. F.; Bergstrom, D.; and Heintz, F. 2024. Bt-GAN: Generating Fair Synthetic Health Data via Bias-Transforming Generative Adversarial Networks. *Journal of Artificial Intelligence Research*, 79: 1313–1341.

Saatchi, Y.; and Wilson, A. G. 2017. Bayesian GAN. In *Advances in Neural Information Processing Systems*.

Spot Intelligence. 2023. Mode Collapse in GANs Explained: How to Detect It and Practical Solutions. Accessed August 29, 2025.

Vuletić, M.; Cucuringu, M.; and Prenzel, F. 2023. Fin-GAN: Forecasting and Classifying Financial Time Series via Generative Adversarial Networks. SSRN preprint.