Generative Uncertainty in Diffusion Models

Metod Jazbec[*1] **Eliot Wong-Toi**[2] **Guoxuan Xia**[3] **Dan Zhang**[4] **Eric Nalisnick**[5] **Stephan Mandt**[2]

[1]UvA-Bosch Delta Lab, University of Amsterdam [2]University of California, Irvine
[3]Imperial College, London [4]Bosch Center for AI [5]Johns Hopkins University

Abstract

Diffusion models have recently driven significant breakthroughs in generative modeling. While state-of-the-art models produce high-quality samples *on average*, individual samples can still be low quality. Detecting such samples without human inspection remains a challenging task. To address this, we propose a Bayesian framework for estimating *generative uncertainty* of synthetic samples. We outline how to make Bayesian inference practical for large, modern generative models and introduce a new semantic likelihood (evaluated in the latent space of a feature extractor) to address the challenges posed by high-dimensional sample spaces. Through our experiments, we demonstrate that the proposed generative uncertainty effectively identifies poor-quality samples and significantly outperforms existing uncertainty-based methods. Notably, our Bayesian framework can be applied *post-hoc* to any pretrained diffusion or flow matching model (via the Laplace approximation), and we propose simple yet effective techniques to minimize its computational overhead during sampling.

1 INTRODUCTION

Diffusion (and flow-matching) models [Sohl-Dickstein et al., 2015, Song et al., 2021a,c, Lipman et al., 2023] have recently pushed the boundaries of generative modeling due to their strong theoretical underpinnings and scalability. Across various domains, they have enabled the generation of increasingly realistic samples [Rombach et al., 2022, Esser et al., 2024, Li et al., 2024]. Despite the impressive progress, state-of-the-art models can still generate low quality images that contain artefacts and fail to align with the provided conditioning information. This poses a challenge for deploying diffusion models, as it can lead to a poor user experience by requiring multiple generations to manually find an artefact-free sample.

Bayesian inference has long been applied to detect poor-quality predictions in predictive models [MacKay, 1992b, Gal et al., 2016, Wilson, 2020, Arbel et al., 2023]. By capturing the uncertainty of the model parameters due to limited training data, each prediction can be assigned a *predictive uncertainty*, which, when high, serves as a warning that the prediction may be unreliable. Despite its widespread use for principled uncertainty quantification in predictive models, Bayesian methodology has been far less commonly applied to detecting poor generations in generative modeling. This raises a key question: *How can Bayesian principles help us detect poor generations?*

In this work, we propose a Bayesian framework for estimating *generative uncertainty* in modern generative models, such as diffusion. To scale Bayesian inference for large diffusion models, we employ the (last-layer) Laplace approximation [MacKay, 1992a, Ritter et al., 2018, Daxberger et al., 2021a]. Additionally, to address the challenge posed by the high-dimensional sample spaces of data such as natural images, we introduce a semantic likelihood, where we leverage pretrained image encoders (such as CLIP [Radford et al., 2021]) to compute variability in a latent, *semantic* space instead. Through our experiments, we demonstrate that generative uncertainty is an effective tool for detecting low-quality samples and propose simple strategies to minimize the sampling overhead introduced by Bayesian inference. In particular, we make the following contributions:

1. We formalize the notion of *generative uncertainty* and propose a method to estimate it for modern generative models (Section 3). Analogous to how predictive uncertainty helps identify unreliable predictions in predictive models, generative uncertainty can be used to detect low-quality generations in generative models.
2. We show that our generative uncertainty strongly outperforms previous uncertainty-based approaches

[*]Corresponding author: <m.jazbec@uva.nl>

Proceedings of the 41st Conference on Uncertainty in Artificial Intelligence (UAI 2025), PMLR 286:1837–1858.

for filtering out poor samples [Kou et al., 2024, De Vita and Belagiannis, 2025] (Section 4.2). Additionally, we achieve competitive performance with non-uncertainty-based methods, such as realism score [Kynkäänniemi et al., 2019] and rarity score [Han et al., 2023], while also highlighting the complementary benefits of uncertainty (Appendix C.5).

3. We propose effective strategies to reduce the sampling overhead of Bayesian uncertainty (Section 4.3) and demonstrate the applicability of our framework beyond diffusion models by applying it to a (latent) flow matching model (Section C.7).

2 BACKGROUND

2.1 GENERATIVE MODELING

Sampling in Generative Models Modern deep generative models like variational autoencoders (VAEs) [Kingma et al., 2014], generative adversarial networks (GANs) [Goodfellow et al., 2014], and diffusion models differ in their exact probabilistic frameworks and training schemes, yet share a common sampling recipe: start with random noise and transform it into a new data sample [Tomczak, 2022]. Specifically, let $\boldsymbol{x} \in \mathcal{X}$ denote a data sample and $\boldsymbol{z} \in \mathcal{Z}$ an initial noise. A new sample is generated by:

$$\boldsymbol{z} \sim p(\mathbf{z}) \,, \quad \hat{\boldsymbol{x}} = g_\theta(\boldsymbol{z}) \,,$$

where $p(\mathbf{z})$ is an initial noise (prior) distribution, typically a standard Gaussian $\mathcal{N}(0, I)$, and $g_\theta : \mathcal{Z} \to \mathcal{X}$ is a generator function with model parameters $\theta \in \mathbb{R}^P$. Here and throughout the paper, we use $\boldsymbol{z}$ (with a slight abuse of notation) to denote the entire randomness involved in the sampling process.[1]

Diffusion Models The primary focus of this work is on diffusion models [Sohl-Dickstein et al., 2015]. These models operate by progressively corrupting data into Gaussian noise and learning to reverse this process. For a data sample $\boldsymbol{x}_0 \sim q(\mathbf{x})$, the forward (noising) process is defined as

$$\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \; \boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$$

where $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$ and $\{\beta_s\}_{s=1}^{T}$, with $\beta_S \in (0,1)$, is a noise schedule chosen such that $\boldsymbol{x}_T \sim \mathcal{N}(0, I)$ (approximately). In the backward process, a denoising network f_θ is learned via a simplified regression objective (among various possible parameterizations, see Song et al. [2021c] or Karras et al. [2022]):

$$\mathcal{L}(\theta) = \mathbb{E}_{t, \boldsymbol{x}_0, \boldsymbol{\epsilon}}\Big[\big|\big| f_\theta(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) - \boldsymbol{\epsilon}\big|\big|_2^2\Big] \,. \quad (1)$$

After training, diffusion models generate new samples via a generator function, $g_{\hat{\theta}}$, which consists of sequentially applying the learned denoiser, $f_{\hat{\theta}}$, and following specific transition rules from samplers such as DDPM [Ho et al., 2020] or DDIM [Song et al., 2021a].

2.2 BAYESIAN DEEP LEARNING

Bayesian neural networks (BNNs) go beyond point predictions and allow for principled uncertainty quantification [Neal, 1995, Kendall and Gal, 2017, Jospin et al., 2022]. Let $h_\psi : \mathcal{X} \to \mathcal{Y}$ denote a predictive model with parameters $\psi \in \mathbb{R}^O$ and $\mathcal{D} = \{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n=1}^N$ denote training data. Instead of finding a single fixed set of parameters, $\hat{\psi} = \arg\max \mathcal{L}(\psi; \mathcal{D})$, that maximizes a chosen objective function $\mathcal{L}$, BNNs specify a prior $p(\psi)$ over model parameters and define a likelihood $p(\mathbf{y}|h_\psi(\boldsymbol{x}))$, which together yield a posterior distribution via Bayes rule: $p(\psi|\mathcal{D}) \propto p(\psi)\prod_{n=1}^N p(\boldsymbol{y}_n|h_\psi(\boldsymbol{x}_n))$. Under this Bayesian view, a predictive model for a new test point $\boldsymbol{x}_*$ is then obtained via the posterior predictive distribution:

$$p(\mathbf{y}|\boldsymbol{x}_*, \mathcal{D}) = \mathbb{E}_{p(\psi|\mathcal{D})}\big[p(\mathbf{y}|h_\psi(\boldsymbol{x}_*))\big].$$

For large models, finding the exact posterior distribution is computationally intractable, hence an approximate posterior $q(\psi|\mathcal{D})$ is used instead. Popular approaches for approximate inference include deep ensembles [Lakshminarayanan et al., 2017, Wilson and Izmailov, 2020], variational inference [Blundell et al., 2015, Zhang et al., 2018], SWAG [Mandt et al., 2017, Maddox et al., 2019], and Laplace approximation [Daxberger et al., 2021a]. Moreover, to alleviate computational overhead, it is common to give a 'Bayesian treatment' only to a subset of parameters [Kristiadi et al., 2020, Daxberger et al., 2021b, Sharma et al., 2023]. Finally, the intractable expectation integral in the posterior predictive is approximated via Monte-Carlo (MC) sampling:

$$p(\mathbf{y}|\boldsymbol{x}_*, \mathcal{D}) \approx \frac{1}{M}\sum_{m=1}^{M} p(\mathbf{y}|h_{\psi_m}(\boldsymbol{x}_*)), \;\; \psi_m \sim q(\psi|\mathcal{D}), \quad (2)$$

with M denoting the number of MC samples. By measuring the variability of the posterior predictive distribution, e.g., its entropy, one can obtain an estimate of the model's predictive uncertainty for a given test point $u(\boldsymbol{x}_*)$. The utility of such uncertainties has been demonstrated on a wide range of tasks such as out-of-distribution (OOD) detection [Daxberger et al., 2021a], active learning [Gal et al., 2017], and detection of influential samples [Nickl et al., 2024].

[1] This distinction matters for diffusion models. In DDIM [Song et al., 2021a] and ODE sampling [Song et al., 2021c], randomness is only present at the start of the sampling process (akin to VAEs and GANs). In contrast, in DDPM [Ho et al., 2020] and SDE sampling [Song et al., 2021c], randomness is introduced at every step throughout the sampling process.

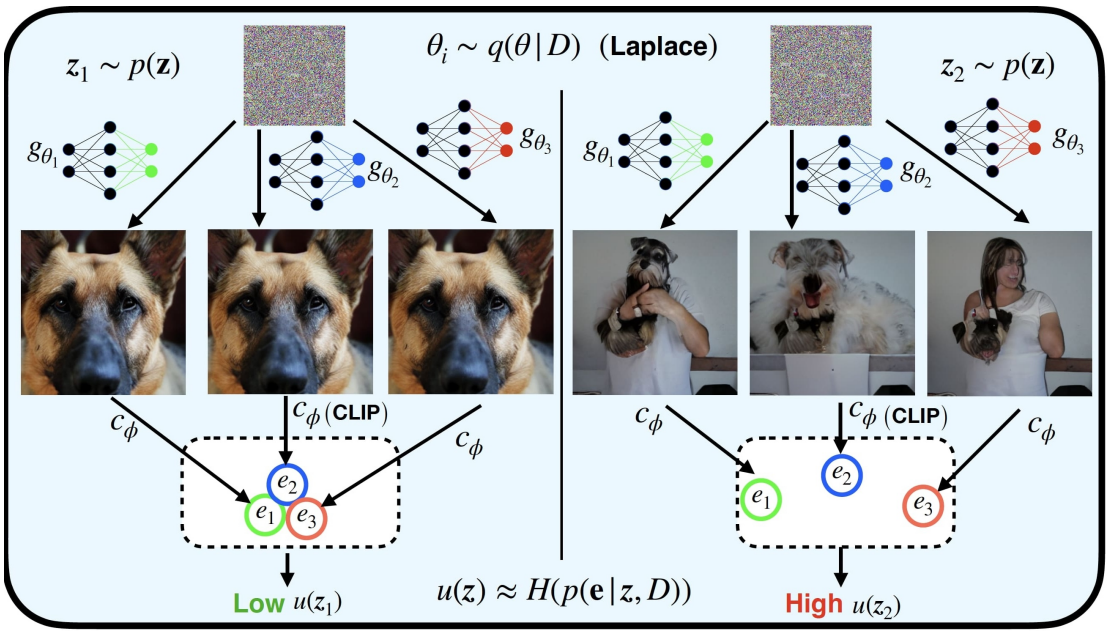

Figure 1: **Generative uncertainty.** Illustration of how we compute generative uncertainty for a fixed random noise $\boldsymbol{z}$. For a diffusion model g_θ, we draw M parameter sets $\{\theta_1, \ldots, \theta_M\}$ from the approximate posterior $q(\theta|D)$ over diffusion model's parameters (here: last-layer Laplace with $M = 3$). Each model g_{θ_m} maps $\boldsymbol{z}$ to an image $\hat{\boldsymbol{x}}_m$. Embedding all $\hat{\boldsymbol{x}}_m$ with a frozen encoder c_ϕ (e.g., CLIP) gives M semantic feature vectors $\boldsymbol{e}_m$; the variability (e.g., entropy) of these vectors is the final uncertainty $u(\boldsymbol{z})$. Low-uncertainty (left) corresponds to consistent, high-quality images, whereas high-uncertainty (right) reveals model disagreement and poor, discordant outputs.

3 GENERATIVE UNCERTAINTY VIA BAYESIAN INFERENCE

While Bayesian neural networks (BNNs) have traditionally been applied to predictive models to estimate *predictive uncertainty*, in this section we demonstrate how to apply them to diffusion to estimate *generative uncertainty* (see Figure 1 and Algorithm 1 for an overview of our method). Later in Section 4, we show that generative uncertainty can be used to detect poor-quality samples. Our focus is on generative models for natural images, where $\boldsymbol{x} \in \mathbb{R}^{H \times W \times C}$. For ease of exposition, we consider unconditional generation in this section, though our methodology can also be applied directly to conditional models (see Section 4.2).

3.1 GENERATIVE UNCERTAINTY

As in traditional Bayesian predictive models (cf. Section 2.2), the central principle for obtaining a Bayesian notion of uncertainty in diffusion models is the posterior predictive distribution:

$$p(\mathbf{x}|\boldsymbol{z}, \mathcal{D}) = \mathbb{E}_{p(\theta|\mathcal{D})}\left[p(\mathbf{x}|g_\theta(\boldsymbol{z}))\right]. \tag{3}$$

Here, as before in Section 2.1 , we use $\boldsymbol{z}$ to denote the entire randomness involved in the diffusion sampling process. Generative uncertainty is then defined as the variability of the posterior predictive:

$$u(\boldsymbol{z}) := \mathcal{V}(p(\mathbf{x}|\boldsymbol{z}, \mathcal{D})) \tag{4}$$

where $\mathcal{V}(\cdot)$ denotes the variability measure, such as entropy. We propose a tractable estimator of the posterior predictive later in Eq. 8.

In the same way that the predictive uncertainty $u(\boldsymbol{x}_*)$, of a predictive model provides insight into the quality of its prediction for a new test point $\boldsymbol{x}_*$, the generative uncertainty $u(\boldsymbol{z})$ of a generative model g_θ should offer information about the quality of the generation $g_\theta(\boldsymbol{z})$ for a 'new' random noise sample $\boldsymbol{z}$. We demonstrate this relationship experimentally in Section 4. Next, we discuss how to make Bayesian inference on (large) diffusion models computationally tractable.

3.2 LAST-LAYER LAPLACE APPROXIMATION

State-of-the-art diffusion models are extremely large (100M to 1B+ parameters) and can take weeks to train. Consequently, the computational overhead of performing Bayesian inference on such large models is of significant concern. For instance, while deep ensembles are a convenient and popular approach for predictive models [Lakshminarayanan et al., 2017], the sheer size of diffusion models renders naive ensembling infeasible. To address this, we adopt the Laplace approximation [MacKay, 1992a, Shun and McCullagh,

1995] to find the approximate posterior $q(\theta|\mathcal{D})$. The Laplace approximation is among the most computationally efficient approximate inference methods while still offering competitive performance [Daxberger et al., 2021a]. Moreover, a particularly appealing feature of the Laplace approximation is that it can be applied *post-hoc* to any diffusion model. We leverage this property in Section 4, where we apply it to a variety of popular diffusion and flow-matching models.

The Laplace approximation of the posterior is given by:

$$q(\theta|\mathcal{D}) = \mathcal{N}(\theta|\hat{\theta}, \Sigma),\ \Sigma = \left(\nabla^2_\theta \mathcal{L}(\theta; \mathcal{D})\big|_{\hat{\theta}}\right)^{-1}, \quad (5)$$

where $\hat{\theta}$ represents the parameters of a pre-trained diffusion model, and Σ is the inverse Hessian of the diffusion training loss from Eq. 1. To reduce the computational cost further, we apply a 'Bayesian' treatment only to the last layer of the denoising network f_θ.

Note that for the Laplace approximation to be theoretically valid, the loss function should correspond to (log-)likelihood and (log-)prior terms. While the diffusion loss can be interpreted as a log-likelihood up to an additive constant—due to its role as a surrogate for the KL divergence between the noising and denoising processes—this interpretation holds only under appropriate weighting of the loss terms across different time steps [Song et al., 2021b]. Consequently, applying the Laplace approximation directly, without such reweighting, is not fully theoretically justified. Despite this, we find in our experiments (Section 4.2) that applying the Laplace approximation directly to the diffusion losses used in practice (without modifying the loss weighting) still yields meaningful uncertainty estimates that align well with the visual quality of generated samples. That said, we believe that better understanding the theoretical requirements for applying approximate Bayesian inference techniques like Laplace in modern generative models like diffusion remains an important direction for future work that could lead to even more informative uncertainty estimates.

It is also worth noting that the use of last-layer Laplace approximation for diffusion models has been previously proposed in BayesDiff [Kou et al., 2024]. While our implementation of the Laplace approximation closely follows theirs, there are significant differences in how we utilize the approximate posterior, $q(\theta|\mathcal{D})$. Specifically, in our approach, we use it within the traditional Bayesian framework (Eq. 3) to sample new diffusion model parameters, leaving the diffusion sampling process, g_θ, unchanged. In contrast, BayesDiff resamples new weights from $q(\theta|\mathcal{D})$ at every diffusion sampling step t, which necessitates substantial modifications to the diffusion sampling process through their *variance propagation* approach. We later demonstrate in Section 4.2 that modifications such as variance propagation are unnecessary for obtaining Bayesian generative uncertainty and staying closer to the traditional Bayesian setting leads to the best empirical performance.

Algorithm 1: Diffusion Sampling with Generative Unc.

Input : random noise $\boldsymbol{z}$, pretrained diffusion model $g_{\hat{\theta}}$, Laplace posterior $q(\theta|\mathcal{D})$ (Eq. 5), number of MC samples M, semantic feature extractor c_ϕ, semantic likelihood noise σ

Output : generated sample $\hat{\boldsymbol{x}}_0$, generative uncertainty estimate $u(\boldsymbol{z})$

Generate a sample $\hat{\boldsymbol{x}}_0 = g_{\hat{\theta}}(\boldsymbol{z})$
Get semantic features $\boldsymbol{e}_0 = c_\phi(\hat{\boldsymbol{x}}_0)$
for $m = 1 \to M$ **do**
  $\theta_m \sim q(\theta|\mathcal{D})$
  $\hat{\boldsymbol{x}}_m = g_{\theta_m}(\boldsymbol{z})$
  $\boldsymbol{e}_m = c_\phi(\hat{\boldsymbol{x}}_m)$
end
Compute $p(\boldsymbol{x}|\boldsymbol{z}, \mathcal{D})$ using $\{\boldsymbol{e}_m\}_{m=0}^M$ (Eq. 8)
Compute the entropy $u(\boldsymbol{z}) = H(p(\boldsymbol{x}|\boldsymbol{z}, \mathcal{D}))$
return $\hat{\boldsymbol{x}}_0, u(\boldsymbol{z})$

3.3 SEMANTIC LIKELIHOOD

We next discuss the choice of likelihood for estimating generative uncertainty in diffusion models. Since the denoising problem in diffusion is modeled as a (multi-output) regression problem, the most straightforward approach is to place a simple Gaussian distribution over the generated sample:

$$p(\mathbf{x}|g_\theta(\boldsymbol{z})) = \mathcal{N}(\mathbf{x} \mid g_\theta(\boldsymbol{z}), \sigma^2 I), \quad (6)$$

where σ^2 represents the observation noise.

However, as we will demonstrate in Section 4, this likelihood leads to non-informative estimates of generative uncertainty (Eq. 4). The primary issue is that the sample space of natural images is high-dimensional (i.e., $|\mathcal{X}| = HWC$). Consequently, placing the likelihood directly in the sample space causes the variability of the posterior predictive distribution to be based on pixel-level differences. This is problematic because it is well-known that two images can appear nearly identical to the human eye while exhibiting a large L_2-norm difference in pixel space $\mathcal{X}$ (see, for example, the literature on adversarial examples [Szegedy et al., 2013]). To get around this, we propose to map the generated samples to a 'semantic' latent space, $\mathcal{S}$, via a pre-trained feature extractor, $c_\phi : \mathcal{X} \to \mathcal{S}$ (e.g., an inception-net [Szegedy et al., 2016] or a CLIP encoder [Radford et al., 2021]). The resulting *semantic likelihood* has the form

$$p(\mathbf{x}|g_\theta(\boldsymbol{z}); \phi) = \mathcal{N}(\mathbf{e}(\mathbf{x}) \mid c_\phi\big(g_\theta(\boldsymbol{z})\big), \sigma^2 I) \quad (7)$$

where $\mathbf{e}(\mathbf{x}) \in \mathcal{S}$ is the (random) vector of semantic features. By combining the (last-layer) Laplace approximate posterior and the semantic likelihood, we can now approximate

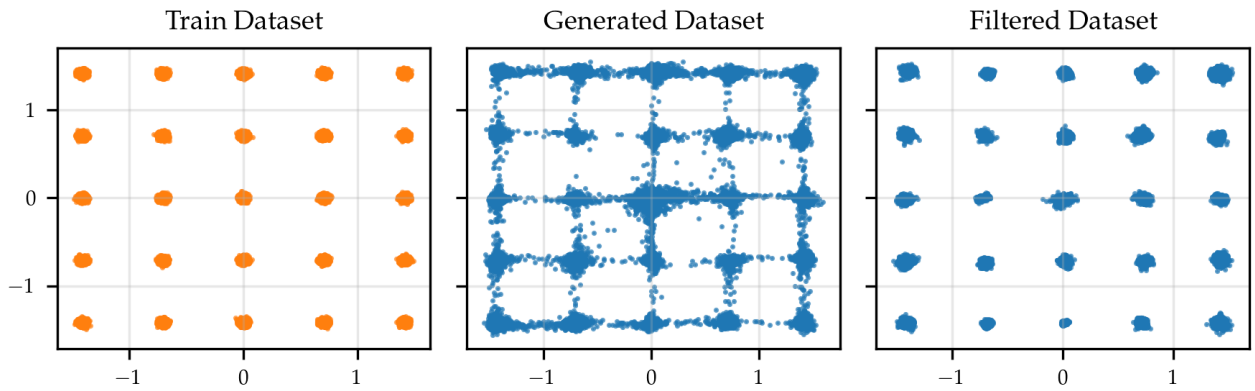

Figure 2: Illustration of how generative uncertainty can be effective for filtering our poor generations on a simple 2D Gaussian dataset. *Left:* training data consisting of 25 separate Gaussian modes. *Middle:* generated samples using a trained diffusion model and a DDPM sampler ($T = 1000$). *Right:* same set of generated samples after removing 50% of generations with the highest estimated generative uncertainty.

the posterior predictive (Eq. 3) as

$$p(\mathbf{x}|\boldsymbol{z}, \mathcal{D}) \approx \mathcal{N}\big(\mathbf{e}(\mathbf{x}) \,\big|\, \bar{e},\, \mathrm{Diag}\big(\frac{1}{M}\sum_{m=1}^{M} e_m^2 - \bar{e}^2\big) + \sigma^2 I\big)\,,$$

$$\bar{e} = \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{e}_m, \quad \boldsymbol{e}_m = c_\phi\big(g_{\theta_m}(\boldsymbol{z})\big),\ \theta_m \sim q(\theta|\mathcal{D}), \quad (8)$$

where M denotes the number of Monte Carlo samples. Additionally, we approximate the posterior predictive with a single Gaussian via moment matching here, a common practice in Bayesian neural networks for regression problems [Lakshminarayanan et al., 2017, Antorán et al., 2020]. A more detailed derivation is provided in Appendix B.1.

Unlike in the posterior predictive for predictive models (Eq. 2), where it is used to obtain both the prediction and the associated uncertainty, the generative posterior predictive (Eq. 8) is used solely to estimate the generative uncertainty $u(\boldsymbol{z})$. The actual samples $\hat{\boldsymbol{x}}$ are still generated using the pre-trained diffusion model $g_{\hat{\theta}}$ (see Algorithm 1). As a variability measure $\mathcal{V}(\cdot)$ in our generative uncertainty framework, we propose to use entropy (denoted with $H(\cdot)$ in Algorithm 1) due to its simplicity and widespread use in quantifying predictive uncertainty. However, we note that alternative measures of variability, such as pairwise-distance estimators (PAiDEs) [Berry and Meger, 2023], can also be employed.

4 EXPERIMENTS

In our experiments, we begin by demonstrating that generative uncertainty serves as an effective method for identifying poor samples in diffusion models, using a simple synthetic dataset (Section 4.1). We then show that our proposed approximations (Sections 3.2 and 3.3) enable the estimation of generative uncertainty in large, modern diffusion models applied to high-dimensional natural images (Section 4.2). Additionally, we discuss the sampling overhead introduced by our Bayesian approach and show that it can be effectively reduced (Section 4.3). Finally, we extend our Bayesian framework beyond diffusion by applying it to detect low-quality samples in a (latent) flow matching model (Appendix C.7). Our code is available at `https://github.com/metodj/DIFF-UQ`.

4.1 TOY DEMONSTRATION

To illustrate the potential of generative uncertainty for detecting poor generations, we adopt the setting from Aithal et al. [2024]. Specifically, we use a 2D synthetic dataset with 25 distinct modes (Figure 2, *left*) to train a small diffusion model. We then generate 50K samples using a DDPM sampler (Figure 2, *middle*). While the generated samples cover the 25 modes well, many 'hallucinated' samples also appear between the modes of the training data. Following Aithal et al. [2024], we consider such samples to be poor generations, as they are highly unlikely under the true data-generating distribution.

Next, we train an ensemble of diffusion models ($M = 5$) and use it to estimate the generative uncertainty of each of the 50K generated samples. We then filter out the 50% of samples with the highest estimated uncertainty and plot the remaining ones in Figure 2, *right*. As shown in the plot, this uncertainty-based filtering effectively removes all poor generations between modes, indicating that generative uncertainty can serve as a reliable indicator of sample quality. Note that in this simple toy setting, we use neither the Laplace approximation (relying instead on a diffusion deep ensemble) nor the semantic likelihood. In the following section, we show how both can be employed to extend generative uncertainty estimation to the more realistic setting of natural images.

4.2 DETECTING LOW-QUALITY GENERATIONS

To demonstrate that our proposed generative uncertainty is effective for detecting low-quality generations also on high-dimensional data such as natural images, we follow the experimental setup from prior work on uncertainty-based filtering [Kou et al., 2024, De Vita and Belagiannis, 2025]. Specifically, we generate 12K samples using a given diffusion model and compute the uncertainty estimate for each sample. We then select $n \in \{6\text{K}, 7\text{K}, \dots, 11\text{K}\}$ samples with the *lowest* uncertainty. If uncertainty reliably reflects the visual quality of generated samples, filtering based on it should yield greater improvements in population-level metrics (such as FID) compared to selecting a random subset of n images.

Implementation Details To ensure a fair comparison with BayesDiff [Kou et al., 2024], we adopt their proposed implementation of the last-layer Laplace approximation. Specifically, we use an Empirical Fisher approximation of the Hessian with a diagonal factorization [Daxberger et al., 2021a]. When computing the posterior predictive distribution (Eq. 8), we use $M = 5$ Monte Carlo samples. For the semantic feature extractor c_ϕ, we leverage a pretrained CLIP encoder [Radford et al., 2021]. We set the observation noise to $\sigma^2 = 0.001$ (Eq. 7). Additional implementation details are provided in Appendix D.

Baselines We compare our proposed generative uncertainty to existing uncertainty-based approaches for detecting low-quality samples: BayesDiff and the aleatoric uncertainty (AU) approach proposed by De Vita and Belagiannis [2025]. BayesDiff estimates epistemic uncertainty in diffusion models using a last-layer Laplace approximation and tracks this uncertainty throughout the entire sampling process. In contrast, in AU, uncertainty is computed by measuring the sensitivity of intermediate diffusion scores to random perturbations. Unlike our approach, both methods estimate uncertainty directly in pixel space.

Evaluation Metrics In addition to the widely used Fréchet Inception Distance (FID) [Heusel et al., 2017] for evaluating the quality of a filtered set of images, we also report *precision* and *recall* metrics [Sajjadi et al., 2018, Kynkäänniemi et al., 2019]. To compute these quantities we fit two manifolds in feature space: one for the generated images and another for the reference (training) images. Precision is the proportion of generated images that lie in the reference image manifold, while recall is the proportion of reference images that lie in the generated image manifold. Precision measures the quality (or fidelity) of generated samples, whereas recall quantifies their diversity (or coverage over the reference distribution).

Results We present our main results on the ImageNet dataset in Figure 3 for UViT model [Bao et al., 2023] and in Figure 7 for ADM model [Dhariwal and Nichol, 2021]. We first observe that existing uncertainty-based approaches (BayesDiff and AU) result in little to no improvement in metrics that assess sample quality (FID and precision). In contrast, our generative uncertainty method leads to significant improvements in terms of both FID and precision. For example, on the UViT model, a subset of $n = 10\text{K}$ images selected based on our uncertainty measure achieves an FID of 7.89, significantly outperforming both the Random baseline (9.45) and existing uncertainty-based methods (BayesDiff 9.16, AU 9.20).

Next, in order to qualitatively demonstrate the effectiveness of our approach, we show 25 samples with the highest and lowest generative uncertainty (out of the original 12K samples) according to our method in Figure 4. High-uncertainty samples exhibit numerous artefacts, and in most cases, it is difficult to determine what exactly they depict. Combined with the quantitative results in Figures 3&7, this supports our hypothesis that (Bayesian) generative uncertainty is an effective metric for identifying low-quality samples. Conversely, the lowest-uncertainty samples are of high quality, with most appearing as 'canonical' examples of their respective (conditioning) class.

For comparison, in Figure 8 we also depict the 25 'worst' and 'best' samples according to the uncertainty estimate from BayesDiff. It is evident that their uncertainty is less informative for sample quality than ours. Moreover, their uncertainty measure appears to be very sensitive to the background pixels. Most images with the highest uncertainty have a 'cluttered' background, whereas most images with the lowest uncertainty have a 'clear' background. We attribute this issue to the fact that in BayesDiff the uncertainty is computed directly in the pixel space, unlike in our approach where we use the semantic likelihood (Section 3.3) to move away from the (high-dimensional) sample space. To further verify the importance of the semantic likelihood, we perform an ablation where we compute the generative uncertainty directly in the pixel-space (see — lines). As seen in Figures 3&7, this leads to worse FID/precision numbers in most cases compared to using the proposed semantic likelihood. Moreover, based on samples in Figure 10, it is clear that without semantic likelihood, our uncertainty becomes overly sensitive to the background pixels in the same way as in BayesDiff.

Lastly, we observe that filtering based on our generative uncertainty results in some loss of sample diversity, as evidenced by lower recall scores (see *right* plots in Figures 3&7). We attribute this to the fact that, in our main experiment, 12K images are generated such that all 1000 ImageNet classes are represented.[2] Since certain classes produce images with higher uncertainty (see Appendix C.6

[2]Following Kou et al. [2024], we use class-conditional diffusion models but randomly sample a class for each of the 12K generated samples.

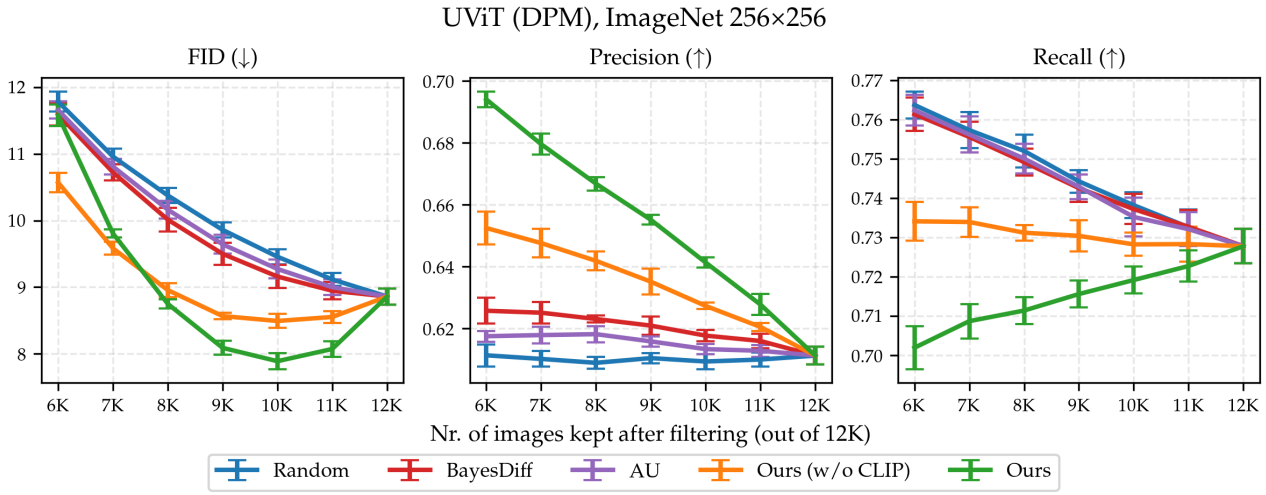

Figure 3: Image generation results for $n \in \{6\text{K}, 7\text{K}, \ldots, 11\text{K}\}$ filtered samples (out of 12K) for UViT diffusion model [Bao et al., 2023]. Our generative uncertainty outperforms previously proposed uncertainty-based approaches (AU [De Vita and Belagiannis, 2025], BayesDiff [Kou et al., 2024]) in terms of image quality, as indicated by higher FID (*left*) and precision (*middle*) scores. We report mean values along with standard deviations over 5 runs with different random seeds.

Figure 4: Images with highest (*left*) and lowest (*right*) generative uncertainty amongst 12K generations using a UViT diffusion model [Bao et al., 2023]. Generative uncertainty correlates with visual quality: high-uncertainty samples exhibit numerous artefacts, whereas low-uncertainty samples resemble canonical images of their respective conditioning class.

for a detailed analysis), filtering based on uncertainty inevitably alters the class distribution among the selected samples. Moreover, the trade-off between improving sample quality (precision) and reducing diversity (recall) has been observed before, see for example the literature on classifier(-free) guidance [Ho and Salimans, 2022].

Comparison with realism and rarity scores Lastly, we compare our proposed method with non-uncertainty-based approaches, such as the *realism* score [Kynkäänniemi et al., 2019] and the *rarity* score [Han et al., 2023]. These metrics work by measuring the distance of a generated sample from the data manifold (derived from a reference dataset) in a semantic space spanned by the inception-net features [Szegedy et al., 2016]. Notably, prior work [Kou et al., 2024, De Vita and Belagiannis, 2025] has not considered such comparisons, which we believe are essential for assessing the practical utility of uncertainty-based filtering.

Figure 5: Images with the highest (*bottom*) and the lowest (*top*) generative uncertainty among 128 generations using a UViT diffusion model for 2 classes: `black swan` (*left*) and `Tibetan terrier` (*right*).

For realism, we retain the n images with the highest scores, whereas for rarity, we keep those with the lowest scores. As shown in Figures 13&14, our generative uncertainty performs on par with realism and rarity in terms of FID. However, compared to our method, realism and rarity scores result in sharper precision-recall trade-offs—yielding larger precision gains at the expense of a greater drop in recall.

Furthermore, Table 1 shows that our score can be effectively combined with realism or rarity scores. Specifically for $n = 10$K, combining our score with realism yields an FID of 7.60 on UViT, compared to 8.26 when combining realism and rarity. We attribute higher benefits from ensembling our score to the fact that, while realism and rarity exhibit a strong negative Spearman correlation (-0.85), our uncertainty measure is less correlated with them (-0.27 with realism, 0.38 with rarity), as shown in Figure 15. Taken together, these results indicate that our uncertainty score captures (somewhat) different desirable properties of images compared to realism and rarity.

4.3 IMPROVING SAMPLING EFFICIENCY

We next examine the sampling costs associated with Bayesian inference in diffusion sampling. As shown in Algorithm 1, obtaining an uncertainty estimate $u(\boldsymbol{z})$ for a generated sample $\hat{\boldsymbol{x}}_0 = g_\theta(\boldsymbol{z})$ requires generating M additional samples, resulting in MT additional network function evaluations (NFEs). For the results presented in Section 4.2, we use $M = 5$ and the default number of sampling steps $T = 50$ (×), leading to an additional 250 NFEs for uncertainty estimation—on top of the 50 NFEs required to generate the original sample. Since this overhead may be prohibitively expensive in certain deployment scenarios, we next explore strategies to reduce the sampling cost associated with our generative uncertainty.

The most straightforward approach is to reduce the number of Monte Carlo samples M. Encouragingly, reducing M to as few as 1 still achieves highly competitive performance (see Figure 6). Further efficiency gains can be achieved by reducing the number of sampling steps T, leveraging the flexibility of diffusion models to adjust T on the fly. Importantly, we lower T only for the additional M samples used for uncertainty assessment while keeping the default T for the original sample $\hat{\boldsymbol{x}}_0$ to ensure that the generation quality is not compromised. Taken together, reducing M and T significantly improves the efficiency of our generative uncertainty. Using the ADM model [Dhariwal and Nichol, 2021], our generative uncertainty method with $M = 1$ and $T = 25$ (●) achieves an FID of 10.36, which still strongly outperforms both the Random (11.31) and BayesDiff (11.20) baselines while requiring only 25 additional NFEs.

5 RELATED WORK

Uncertainty quantification in diffusion models has recently gained significant attention. Most related to our work are BayesDiff [Kou et al., 2024], which uses a Laplace approximation to track epistemic uncertainty throughout the sampling process, and De Vita and Belagiannis [2025], which captures aleatoric uncertainty via the sensitivity of diffusion score estimates. Our work extends both by introducing an uncertainty framework that is more general (applicable beyond diffusion), simpler (requiring no sampling modifications), and more effective (see Section 4.2).

Also related is DECU [Berry et al., 2024], which employs an efficient variant of deep ensembles [Lakshminarayanan et al., 2017] to capture the epistemic uncertainty of conditional diffusion models. However, DECU does not consider using uncertainty to detect poor-quality generations, as its framework provides uncertainty estimates at the level of the conditioning variable, whereas ours estimates uncertainty at the level of initial random noise. Similarly, in Chan et al. [2024] the use of hyper-ensembles is proposed to capture epistemic uncertainty in diffusion models for inverse problems such as super-resolution, but, as in DECU, their approach does not provide uncertainty estimates in unconditional settings or in conditional settings with low-dimensional conditioning (such as class-conditional generation). Moreover, both DECU [Berry et al., 2024] and Chan et al. [2024] require modifying and retraining diffusion model components, whereas our approach oper-

ates *post-hoc* with any pretrained diffusion model via the Laplace approximation [Daxberger et al., 2021a]. A recent approach, PUNC [Franchi et al., 2025], focuses only on text-to-image models. The uncertainty of image generation with respect to text conditioning is measured through the alignment between a caption generated from a generated image and the original prompt used to generate said image.

Additionally, a large body of work explores conformal prediction for uncertainty quantification in diffusion models [Angelopoulos et al., 2022, Sankaranarayanan et al., 2022, Teneggi et al., 2023, Belhasin et al., 2023]. However, these approaches are primarily designed for inverse problems (e.g., deblurring), and cannot be directly applied to detect low-quality samples in unconditional generation.

Bayesian inference in generative models has been explored previously outside the domain of diffusion models. Prominent examples include Saatci and Wilson [2017] where a Bayesian version of a GAN is proposed, showing improvements for semi-supervised learning, and Daxberger and Hernández-Lobato [2019], where a Bayesian VAE [Tran et al., 2023] is shown to provide more informative likelihood estimates for the unsupervised out-of-distribution detection compared to the non-Bayesian counterparts [Nalisnick et al., 2019]. Since diffusion models can be interpreted as neural ODEs [Song et al., 2021c], another relevant work is Ott et al. [2023], which employs a Laplace approximation to quantify uncertainty when solving neural ODEs [Chen et al., 2018]. However, Ott et al. [2023] focuses solely on low-dimensional regression problems.

Non-uncertainty based approaches for filtering out poor generations include the realism [Kynkäänniemi et al., 2019], rarity [Han et al., 2023], and anomaly scores [Hwang et al., 2024]. Our work is the first to establish a connection between these scores and uncertainty-based methods, which we hope will inspire the development of even better sample-level metrics in the future. Additionally, a large body of work focuses on specially designed sample-quality scoring models [Gu et al., 2020, Zhao et al., 2024] or, alternatively, on leveraging large pretrained vision-language models (VLMs) [Zhang et al., 2025] for scoring generated images. However, these approaches require either access to sample-quality labels or rely on (expensive) external VLMs. In contrast, our uncertainty-based method requires neither, making it a more accessible and scalable alternative.

6 LIMITATIONS

While we have demonstrated in Section 4 that semantic likelihood is essential for addressing the over-sensitivity of prior work to background pixels [Kou et al., 2024], our reliance on a pretrained image encoder like CLIP [Radford et al., 2021] limits the applicability of our diffusion uncertainty framework to natural images. Removing the dependence on

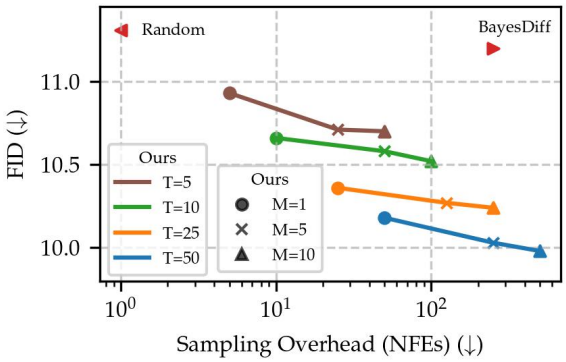

Figure 6: FID results for $n = 10$K ImageNet filtered images using our generative uncertainty on ADM model [Dhariwal and Nichol, 2021]. We vary the number of Monte Carlo samples M and diffusion sampling steps T (see Algorithm 1). By default, we use M=5 with T=50 (×), incurring an additional 250 NFEs for uncertainty estimation. Encouragingly, setting M=1 and T=25 (●) still achieves competitive performance while reducing the sampling overhead by 10x. Lower left is best: better FID and greater computational efficiency.

such encoders would unlock the application our Bayesian framework to other modalities where diffusion models are used, such as molecules [Hoogeboom et al., 2022, Cornet et al., 2024] or text [Gong et al., 2023, Yi et al., 2024]. Exploring whether insights from the literature on uncovering semantic features in diffusion models [Kwon et al., 2023, Luo et al., 2024, Namekata et al., 2024] could help achieve this represents a promising direction for future work.

Moreover, the large size of modern diffusion models necessitates the use of cheap and scalable Bayesian approximate inference techniques, such as the (diagonal) last-layer Laplace approximation employed in our work (following [Kou et al., 2024]). A more comprehensive comparison of available approximate inference methods could be valuable, as improving the quality of the posterior approximation may further enhance the detection of low-quality samples based on Bayesian generative uncertainty.

7 CONCLUSION

We introduced generative uncertainty and demonstrated how to estimate it in modern generative models such as diffusion. Our experiments showed the effectiveness of generative uncertainty in filtering out low-quality samples. For future work, it would be interesting to explore broader applications of Bayesian principles in generative modeling beyond detecting poor-quality generations. Promising directions include guiding synthetic data generation and detecting memorized samples. It would also be worthwhile to further investigate the connection between uncertainty-based filtering and classifier(-free) guidance [Ho and Salimans, 2022], as both exhibit similar precision-recall trade-offs.

Acknowledgements

We thank our reviewers for their thoughtful feedback. This project was generously supported by the Bosch Center for Artificial Intelligence. E.N. did not utilize resources from Johns Hopkins University for this project. S.M. acknowledges funding from the National Science Foundation (NSF) through an NSF CAREER Award IIS-2047418, IIS-2007719, the NSF LEAP Center, the IARPA WRIVA program, and the Hasso Plattner Research Center at UCI. See Appendix for additional disclaimers.

References

Sumukh K Aithal, Pratyush Maini, Zachary Lipton, and J Zico Kolter. Understanding hallucinations in diffusion models through mode interpolation. In *NeurIPS*, 2024.

Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv:2303.08797*, 2023.

Anastasios N Angelopoulos, Amit Pal Kohli, Stephen Bates, Michael Jordan, Jitendra Malik, Thayer Alshaabi, Srigokul Upadhyayula, and Yaniv Romano. Image-to-image regression with distribution-free uncertainty quantification and applications in imaging. In *ICML*, 2022.

Javier Antorán, James Allingham, and José Miguel Hernández-Lobato. Depth uncertainty in neural networks. In *NeurIPS*, 2020.

Julyan Arbel, Konstantinos Pitas, Mariia Vladimirova, and Vincent Fortuin. A primer on bayesian neural networks: review and debates. *arXiv:2309.16314*, 2023.

Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. In *CVPR*, 2023.

Omer Belhasin, Yaniv Romano, Daniel Freedman, Ehud Rivlin, and Michael Elad. Principal uncertainty quantification with spatial correlation for image restoration problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Lucas Berry and David Meger. Escaping the sample trap: Fast and accurate epistemic uncertainty estimation with pairwise-distance estimators. *arXiv:2308.13498*, 2023.

Lucas Berry, Axel Brando, and David Meger. Shedding light on large generative networks: Estimating epistemic uncertainty in diffusion models. In *UAI*, 2024.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *ICML*, 2015.

Matthew Albert Chan, Maria J Molina, and Christopher Metzler. Estimating epistemic and aleatoric uncertainty with a single model. In *NeurIPS*, 2024.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *NeurIPS*, 2018.

François Cornet, Grigory Bartosh, Mikkel N Schmidt, and Christian A Naesseth. Equivariant neural diffusion for molecule generation. In *NeurIPS*, 2024.

Quan Dao, Hao Phung, Binh Nguyen, and Anh Tran. Flow matching in latent space. *arXiv:2307.08698*, 2023.

Erik Daxberger and José Miguel Hernández-Lobato. Bayesian variational autoencoders for unsupervised out-of-distribution detection. *arXiv:1912.05651*, 2019.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux-effortless bayesian deep learning. In *NeurIPS*, 2021a.

Erik Daxberger, Eric Nalisnick, James U Allingham, Javier Antorán, and José Miguel Hernández-Lobato. Bayesian deep learning via subnetwork inference. In *ICML*, 2021b.

Michele De Vita and Vasileios Belagiannis. Diffusion model guided sampling with pixel-wise aleatoric uncertainty estimation. In *WACV*, 2025.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024.

Gianni Franchi, Nacim Belkhir, Dat Nguyen Trong, Guoxuan Xia, and Andrea Pilzer. Towards understanding and quantifying uncertainty for text-to-image generation. In *CVPR*, 2025.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *ICML*, 2017.

Yarin Gal et al. Uncertainty in deep learning. 2016.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and LingPeng Kong. Diffuseq: Sequence to sequence text generation with diffusion models. In *ICLR*, 2023.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014.

Shuyang Gu, Jianmin Bao, Dong Chen, and Fang Wen. Giqa: Generated image quality assessment. In *ECCV*, 2020.

Jiyeon Han, Hwanil Choi, Yunjey Choi, Junho Kim, Jung-Woo Ha, and Jaesik Choi. Rarity score : A new metric to evaluate the uncommonness of synthesized images. In *ICLR*, 2023.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

Emiel Hoogeboom, Vıctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *ICML*, 2022.

Jaehui Hwang, Junghyuk Lee, and Jong-Seok Lee. Anomaly score: Evaluating generative models and individual generated images based on complexity and vulnerability. In *CVPR*, 2024.

Alexander Immer, Matthias Bauer, Vincent Fortuin, Gunnar Rätsch, and Khan Mohammad Emtiyaz. Scalable marginal likelihood estimation for model selection in deep learning. In *ICML*, 2021.

Laurent Valentin Jospin, Hamid Laga, Farid Boussaid, Wray Buntine, and Mohammed Bennamoun. Hands-on bayesian neural networks—a tutorial for deep learning users. *IEEE Computational Intelligence Magazine*, 2022.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *NeurIPS*, 2022.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NeurIPS*, 2017.

Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes. In *ICLR*, 2014.

Siqi Kou, Lei Gan, Dequan Wang, Chongxuan Li, and Zhijie Deng. Bayesdiff: Estimating pixel-wise uncertainty in diffusion via bayesian inference. In *ICLR*, 2024.

Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Being bayesian, even just a bit, fixes overconfidence in relu networks. In *ICML*, 2020.

Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. In *ICLR*, 2023.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *NeurIPS*, 2019.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, 2017.

Zehui Li, Yuhao Ni, Guoxuan Xia, William Beardall, Akashaditya Das, Guy-Bart Stan, and Yiren Zhao. Absorb & escape: Overcoming single model limitations in generating heterogeneous genomic sequences. In *NeurIPS*, 2024.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *ICLR*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. In *NeurIPS*, 2022.

Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. In *NeurIPS*, 2024.

David JC MacKay. Bayesian interpolation. *Neural computation*, 1992a.

David JC MacKay. Information-based objective functions for active data selection. *Neural computation*, 1992b.

Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. In *NeurIPS*, 2019.

Stephan Mandt, Matthew D Hoffman, David M Blei, et al. Stochastic gradient descent as approximate bayesian inference. *Journal of Machine Learning Research*, 2017.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *ICLR*, 2019.

Koichi Namekata, Amirmojtaba Sabour, Sanja Fidler, and Seung Wook Kim. Emerdiff: Emerging pixel-level semantic knowledge in diffusion models. In *ICLR*, 2024.

Radford M Neal. *Bayesian learning for neural networks*. Springer Science & Business Media, 1995.

Peter Nickl, Lu Xu, Dharmesh Tailor, Thomas Möllenhoff, and Mohammad Emtiyaz E Khan. The memory-perturbation equation: Understanding model's sensitivity to data. In *NeurIPS*, 2024.

Katharina Ott, Michael Tiemann, and Philipp Hennig. Uncertainty and structure in neural ordinary differential equations. *arXiv:2305.13290*, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Hippolyt Ritter, Aleksandar Botev, and David Barber. A scalable laplace approximation for neural networks. In *ICLR*, 2018.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Yunus Saatci and Andrew G Wilson. Bayesian gan. In *NeurIPS*, 2017.

Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. In *NeurIPS*, 2018.

Swami Sankaranarayanan, Anastasios Angelopoulos, Stephen Bates, Yaniv Romano, and Phillip Isola. Semantic uncertainty intervals for disentangled latent spaces. In *NeurIPS*, 2022.

Mrinank Sharma, Sebastian Farquhar, Eric Nalisnick, and Tom Rainforth. Do bayesian neural networks need to be fully stochastic? In *AISTATS*, 2023.

Zhenming Shun and Peter McCullagh. Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 1995.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021a.

Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. In *NeurIPS*, 2021b.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021c.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv:1312.6199*, 2013.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.

Jacopo Teneggi, Matthew Tivnan, Web Stayman, and Jeremias Sulam. How to trust your diffusion model: A convex optimization approach to conformal risk control. In *ICML*, 2023.

Lucas Theis. What makes an image realistic? In *ICML*, 2024.

Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models. In *ICLR*, 2016.

Jakub M. Tomczak. *Deep Generative Modeling*. Springer, 2022.

Ba-Hien Tran, Babak Shahbaba, Stephan Mandt, and Maurizio Filippone. Fully bayesian autoencoders with latent sparse gaussian processes. In *ICML*, 2023.

Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *ICML*, 2020.

Andrew G Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. In *NeurIPS*, 2020.

Andrew Gordon Wilson. The case for bayesian deep learning. *arXiv:2001.10995*, 2020.

Qiuhua Yi, Xiangfan Chen, Chenwei Zhang, Zehai Zhou, Linan Zhu, and Xiangjie Kong. Diffusion models in text generation: a survey. *PeerJ Computer Science*, 2024.

Cheng Zhang, Judith Bütepage, Hedvig Kjellström, and Stephan Mandt. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

Zicheng Zhang, Haoning Wu, Chunyi Li, Yingjie Zhou, Wei Sun, Xiongkuo Min, Zijian Chen, Xiaohong Liu, Weisi Lin, and Guangtao Zhai. A-bench: Are lmms masters at evaluating ai-generated images? In *ICLR*, 2025.

Ganning Zhao, Vasileios Magoulianitis, Suya You, and C-C Jay Kuo. A lightweight generalizable evaluation and enhancement framework for generative models and generated samples. In *WACV*, 2024.

APPENDIX

The supplementary material is organized as follows:

- In Appendix A, we provide additional figures.
- In Appendix B.1, we provide more details on our approximation of generative uncertainty (Eq. 8).
- In Appendix C.1, we qualitatively compare our method with BayesDiff [Kou et al., 2024].
- In Appendix C.2, we perform qualitative ablations on our semantic likelihood (Section 3.3).
- In Appendix C.3, we demonstrate how to use our generative uncertainty for pixel-wise uncertainty.
- In Appendix C.4, we show that diffusion's own likelihood is not useful for filtering out poor samples.
- In Appendix C.5, we compare our generative uncertainty to realism [Kynkäänniemi et al., 2019] and rarity [Han et al., 2023] scores.
- In Appendix C.6, we investigate the drop in sample diversity by looking at the average generative uncertainty per conditioning class.
- In Appendix C.7, we apply our generative uncertainty to detect low-quality samples in a latent flow matching model [Dao et al., 2023].
- In Appendix D, we provide implementation and experimental details.

A ADDITIONAL FIGURES

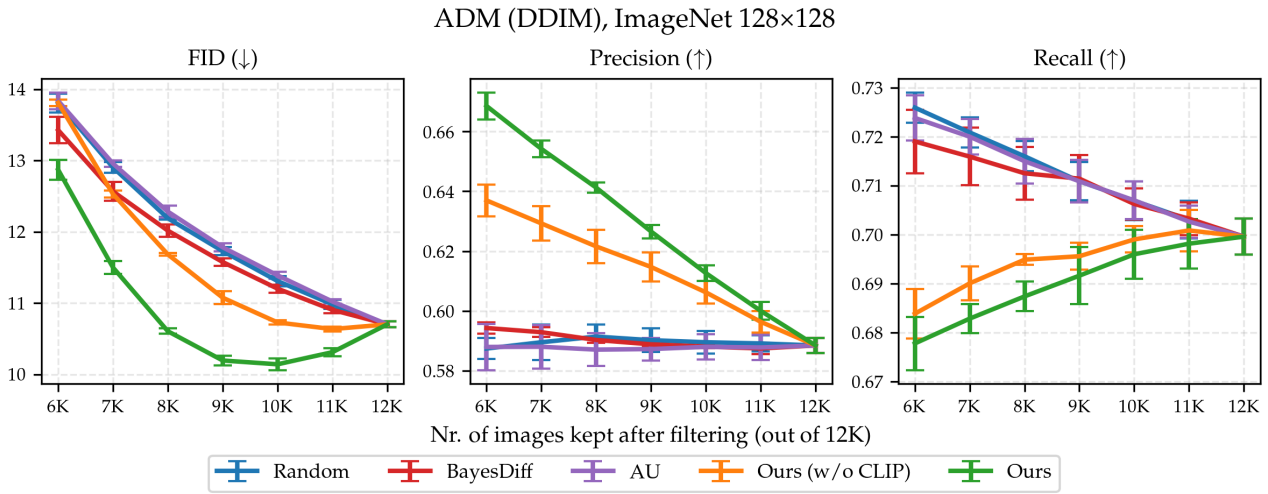

Figure 7: Image generation results for $n \in \{6\text{K}, 7\text{K}, \ldots, 11\text{K}\}$ filtered samples (out of 12K) for ADM diffusion model [Dhariwal and Nichol, 2021]. Our generative uncertainty outperforms previously proposed uncertainty-based approaches (AU [De Vita and Belagiannis, 2025], BayesDiff [Kou et al., 2024]) in terms of image quality, as indicated by higher FID (*left*) and precision (*middle*) scores. We report mean values along with standard deviations over 5 runs with different random seeds.

B DERIVATIONS

B.1 GENERATIVE UNCERTAINTY APPROXIMATION

To ensure maximal clarity of exposition, we first derive a tractable estimator of $u(\boldsymbol{z})$ (Eq. 4) based on 'pixel-space' likelihood (Eq. 6):

$$p(\mathbf{x}|\boldsymbol{z},\mathcal{D}) \overset{Eq.3}{=} \mathbb{E}_{p(\theta|\mathcal{D})}\big[p(\mathbf{x}|g_\theta(\boldsymbol{z})\big] \overset{(1)}{\approx} \frac{1}{M}\sum_{m=1}^{M} p(\mathbf{x}|g_{\theta_m}(\boldsymbol{z})) \overset{Eq.6}{=} \frac{1}{M}\sum_{m=1}^{M} \mathcal{N}(\mathbf{x}|g_{\theta_m}(\boldsymbol{z}),\sigma^2 I) \overset{(2)}{\approx}$$
$$\mathcal{N}\big(\mathbf{x}|\bar{\boldsymbol{g}},\text{Diag}\big(\frac{1}{M}\sum_{m=1}^{M} g_{\theta_m}(\boldsymbol{z})^2 - \bar{\boldsymbol{g}}^2\big) + \sigma^2 I\big) =: q_{\text{pixel}}(\boldsymbol{z})$$

where $\theta_m \sim q(\theta|\mathcal{D})$ (Eq. 5) and $\bar{\boldsymbol{g}} = \frac{1}{M}\sum_{m=1}^{M} g_{\theta_m}(\boldsymbol{z})$. Note that in step (1) we make use of the usual Monte Carlo (MC) approximation of the Bayesian posterior predictive (Eq. 2) using M samples and in step (2) we approximate a mixture of Gaussian with a single Gaussian using moment-matching. Moreover, we consider only the diagonal of the resulting covariance which we do for efficiency reasons (e.g., for ImageNet 256x256 the full covariance has $\sim 4 \cdot 10^{10}$ parameters rendering the diagonal approximation necessary).

A pixel-space generative uncertainty is then obtained by taking the entropy of the resulting Gaussian distribution: $u(\boldsymbol{z}) \approx H(q_{\text{pixel}}(\boldsymbol{z}))$. However, as we show qualitatively in Appendix C.2, uncertainty based on pixel-space likelihood is not particularly informative about the visual quality of the samples as it is overly sensitive to the background pixels—images with simple backgrounds exhibit low uncertainty, whereas images with 'cluttered' background exhibit high uncertainty. This motivates our use of 'semantic-likelihood' (Eq. 7) to arrive at the Eq. 8:

$$p(\mathbf{x}|\boldsymbol{z},\mathcal{D}) \overset{(1)}{\approx} \frac{1}{M}\sum_{m=1}^{M} p(\mathbf{x}|g_{\theta_m}(\boldsymbol{z});\phi) \overset{\text{Eq. 7}}{=} \frac{1}{M}\sum_{m=1}^{M} \mathcal{N}(\boldsymbol{e}(\mathbf{x})|c_\phi(g_{\theta_m}(\boldsymbol{z})),\sigma^2 I) \overset{(2)}{\approx}$$
$$\mathcal{N}\Big(\boldsymbol{e}(\mathbf{x})\Big|\bar{\boldsymbol{e}},\text{Diag}\Big(\frac{1}{M}\sum_{m=1}^{M} \boldsymbol{e}_m^2 - \bar{\boldsymbol{e}}^2\Big) + \sigma^2 I\Big) =: q_{\text{semantic}}(\boldsymbol{z})$$

where $\bar{\boldsymbol{e}} = \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{e}_m,\ \boldsymbol{e}_m = c_\phi(g_{\theta_m}(\boldsymbol{z})),\ \theta_m \sim q(\theta|\mathcal{D})$ and c_ϕ is a pre-trained feature extractor of choice (e.g., CLIP). With step (1) we again denote the MC approximation (Eq. 2) and step (2) denotes moment-matching (with a diagonal covariance approximation). While the resulting posterior predictive based on the semantic likelihood can not be used to generate samples (since the likelihood is over CLIP features $\mathbf{e}(\mathbf{x}) \in \mathcal{S}$ and not data $\mathbf{x} \in \mathcal{X}$), we can still compute its entropy which we use as our final estimate of generative uncertainty $u(\boldsymbol{z}) \approx H(q_{\text{semantic}}(\boldsymbol{z}))$. As described in Section 3.3 and in Algorithm 1, we first generate a sample using a pretrained model $g_{\hat{\theta}}$ and then use the semantic posterior predictive q_{semantic} solely for uncertainty estimation.

At this point it is important to acknowledge that by changing the likelihood to the semantic one we depart from the traditional Bayesian framework where the same likelihood is used both for finding the posterior $q(\theta|\mathcal{D})$ as well as in the approximation of the posterior predictive. However, we would like to emphasize that image generation using modern diffusion models poses specific challenges, which to the best of our knowledge, have not been addressed within the Bayesian framework yet. One such challenge is due to (extremely) high-dimensional sample spaces. For example, in the case of ImageNet 256x256 the dimensionality is $\sim 2 \cdot 10^5$. Our use of feature extractor c_ϕ via the semantic likelihood reduces the dimensionality (down to 512), potentially making the MC approximation using few samples (M) 'easier'. Another challenge is that using a larger number of MC samples is computational prohibitive, since every additional sample corresponds to generating a new sample with a diffusion model which is costly.

We hope that our promising experimental results based on the semantic likelihood (using a few MC samples only) will encourage the Bayesian community to further investigate the choice of the suitable likelihood in high-dimensional spaces (such as those of natural images) and fill-in the potentially missing theoretical gaps (e.g., due to changing the likelihood 'post-hoc').

C ADDITIONAL RESULTS

C.1 QUALITATIVE COMPARISON WITH BAYESDIFF

To further highlight the differences between our generative uncertainty and BayesDiff [Kou et al., 2024], we present samples with the highest and lowest uncertainty according to BayesDiff in Figure 8. These samples are drawn from the same set of 12K ImageNet images generated using the UViT model [Bao et al., 2023] as in Figure 4. Notably, BayesDiff's uncertainty score appears highly sensitive to background pixels—images with high uncertainty tend to have cluttered backgrounds, while those with low uncertainty typically feature clear backgrounds. Furthermore, as reflected in BayesDiff's poor performance in terms of FID and precision (see Figures 3&7), some low-uncertainty examples exhibit noticeable artefacts, whereas certain high-uncertainty samples are of rather high-quality. For example, the image of a dog in the bottom-right corner of the high-uncertainty grid in Figure 8 looks quite good despite being assigned (very) high uncertainty.

Similarly, in Figure 9, we show low- and high-uncertainty samples according to BayesDiff for the same set of 128 images per class as in Figure 5. Once again, we observe that BayesDiff's uncertainty metric is less informative regarding a sample's visual quality compared to our generative uncertainty.

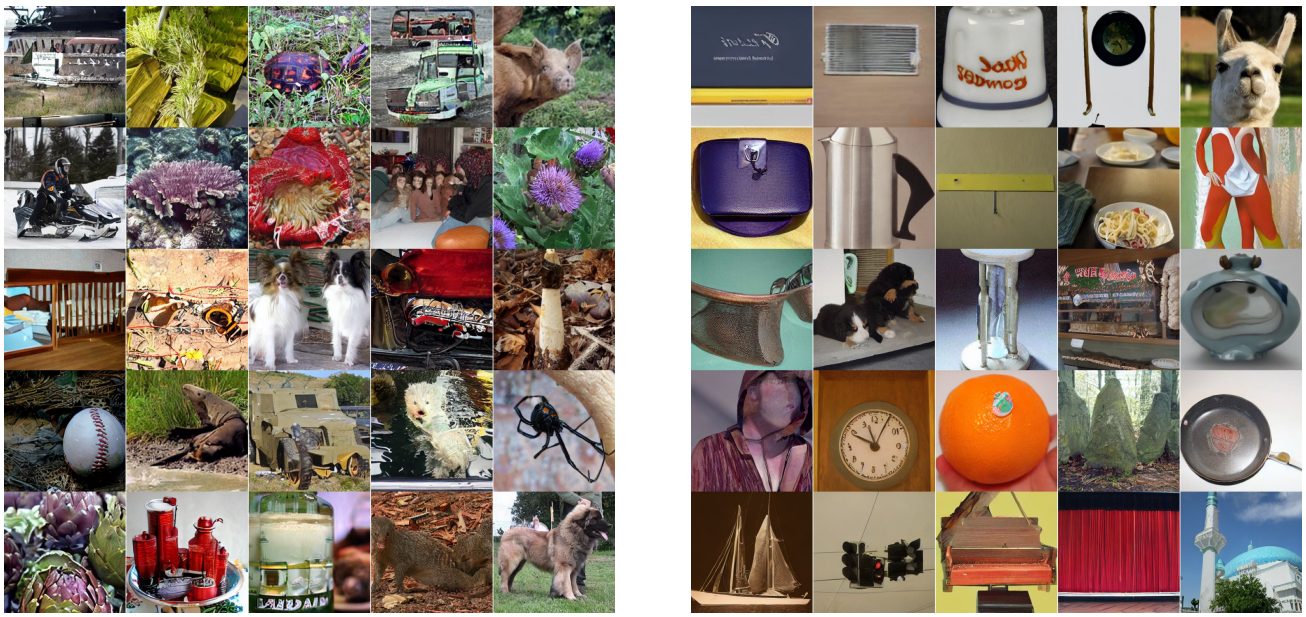

Figure 8: Images with the highest (*left*) and the lowest (*right*) BayesDiff uncertainty among 12K generations using a UViT diffusion model [Bao et al., 2023]. BayesDiff uncertainty correlates poorly with visual quality and is overly sensitive to the background pixels. Same set of 12K generated images is used as in Figure 4 to ensure a fair comparison.

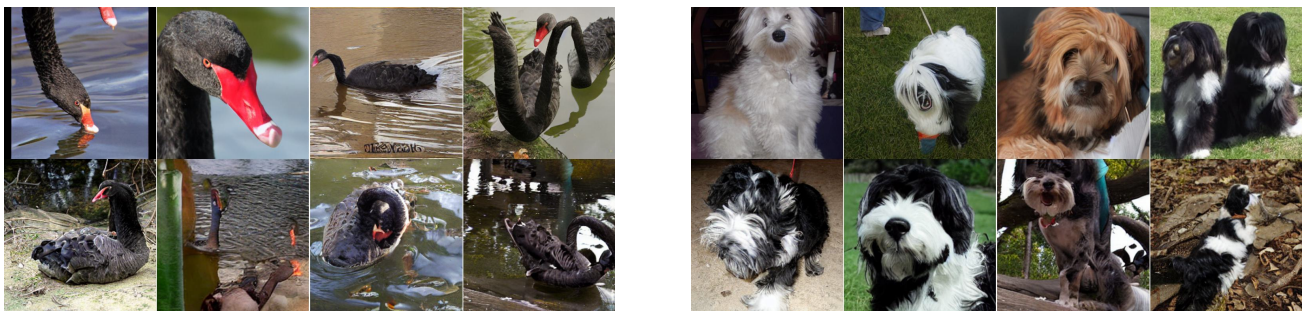

Figure 9: Images with the highest (*bottom*) and the lowest (*top*) BayesDiff uncertainty among 128 generations using a UViT diffusion model for 2 classes: `black swan` (*left*) and `Tibetan terrier` (*right*). Same set of 128 generated images per class is used as in Figure 5 to ensure a fair comparison.

C.2 ABLATION ON SEMANTIC LIKELIHOOD

To highlight the importance of using a semantic likelihood (Section 3.3) when leveraging uncertainty to detect low-quality generations, we conduct an ablation study in which we replace it with a standard Gaussian likelihood applied directly in pixel space (Eq. 6). Figure 10 presents the highest and lowest uncertainty images according to this 'pixel-space' generative uncertainty. Notably, pixel-space uncertainty is overly sensitive to background pixels, mirroring the issue observed in BayesDiff (see Appendix C.1). This highlights the necessity of using semantic likelihood to obtain uncertainty estimates that are truly informative about the visual quality of generated samples.

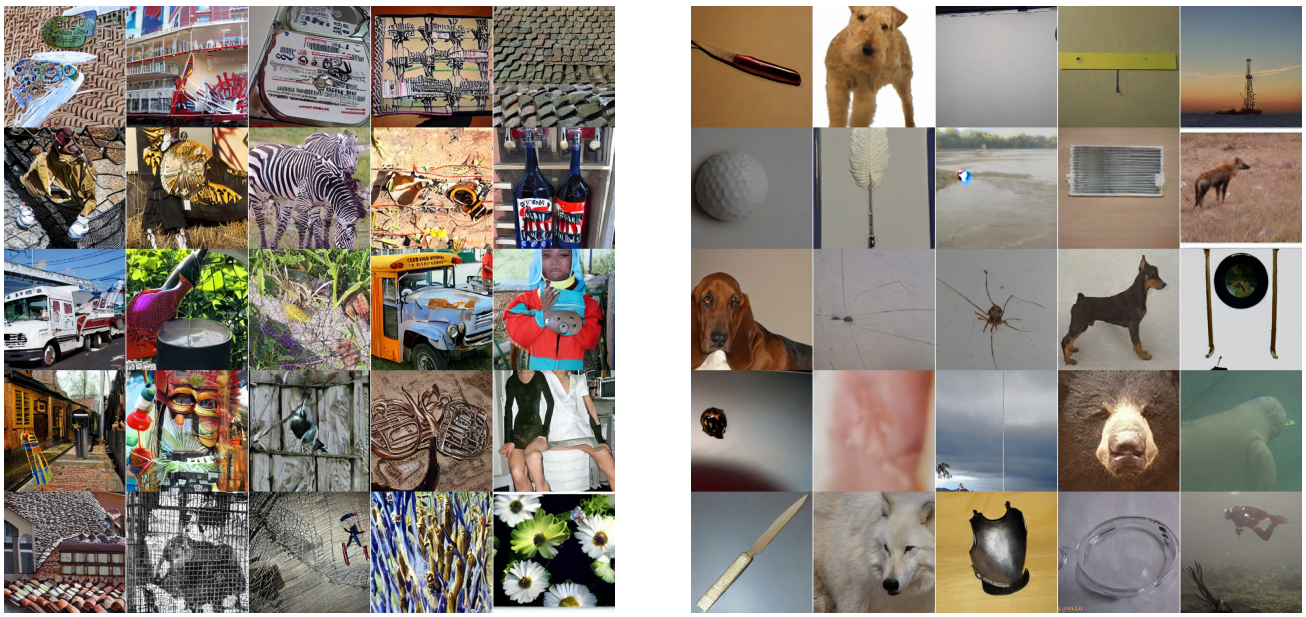

Figure 10: Images with the highest (*left*) and the lowest (*right*) 'pixel-space' generative uncertainty among 12K generations using a UViT diffusion model. Pixel-space uncertainty correlates poorly with visual quality and is overly sensitive to the background pixels. Same set of 12K generated images is used as in Figure 4 to ensure a fair comparison.

C.3 PIXEL-WISE UNCERTAINTY

While not the primary focus of our work, we demonstrate how our generative uncertainty framework (Algorithm 1) can be adapted to obtain pixel-wise uncertainty estimates. This is achieved by replacing our proposed semantic likelihood (Eq. 7) with a standard 'pixel-space' likelihood (Eq. 6). Figure 11 illustrates pixel-wise uncertainty estimates for 5 generated samples.

Although pixel-wise uncertainty received significant attention in past work [Kou et al., 2024, Chan et al., 2024, De Vita and Belagiannis, 2025], there is currently no principled method for evaluating its quality. Most existing approaches rely on qualitative inspection, visualizing pixel-wise uncertainty for a few generated samples (as we do in Figure 11). This further motivates our focus on sample-wise uncertainty estimates, where more rigorous evaluation frameworks—such as improvements in FID and precision on a set of filtered images—enable more meaningful comparisons between different approaches.

C.4 COMPARISON WITH LIKELIHOOD

We compare our generative uncertainty filtering criterion with a likelihood selection approach on the 12K images generated by ADM trained on ImageNet 128x128. Here retain the $n = 10$K generated images with highest likelihood. We utilize the implementation in Dhariwal and Nichol [2021] to compute the bits-per-dimension of each sample (one-to-one with likelihood). The 25 samples with lowest and highest likelihood are shown in Figure 12. Visually, the likelihood objective heavily prefers simple images with clean backgrounds and not necessarily image quality. Note that this is consistent with other works that have reported likelihood to be an inconsistent identifier of image quality [Theis et al., 2016, Theis, 2024]. Quantitative results for image quality were consistent with our qualitative observations. The FID, precision, and recall for the best 10K images according to bits-per-dimension were 11.86 ± 0.0026, 58.23 ± 0.02160, and 70.45 ± 0.0237 over three runs. By point estimate,

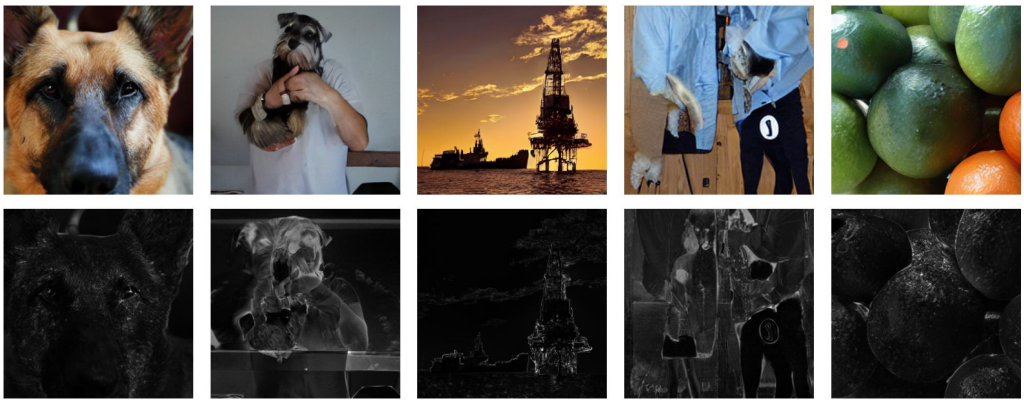

Figure 11: Pixel-wise uncertainty based on our generative uncertainty for 5 generated samples using UViT diffusion.

all three metrics are worse or indistinguishable from the Random baseline ($11.31 \pm 0.07, 58.90 \pm 0.36, 70.68 \pm 0.38$).

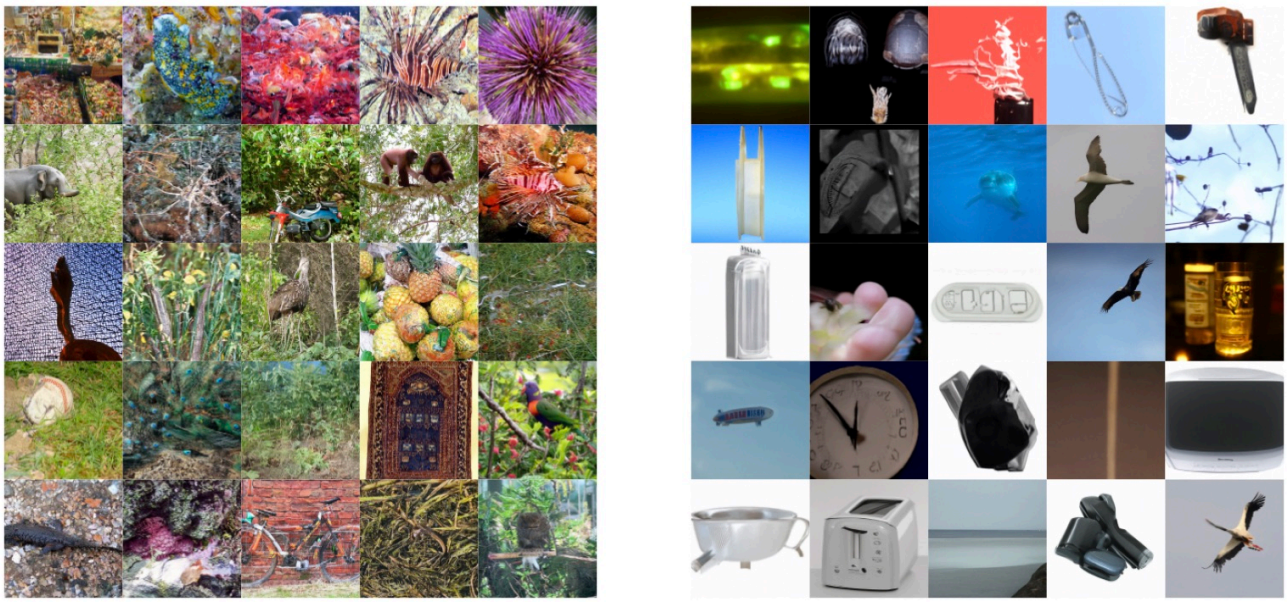

Figure 12: The 25 'worst' (*left*) and 'best' (*right*) samples generated by ADM trained on ImageNet 128x128 selected by lowest and highest likelihood among 12K generations.

C.5 COMPARISON WITH REALISM & RARITY

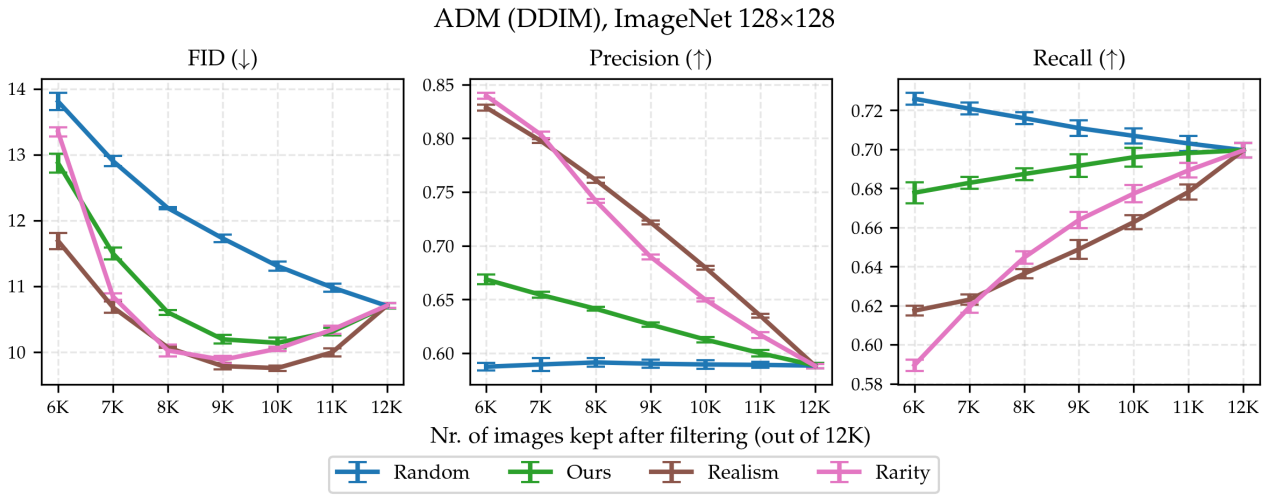

Figure 13: Image generation results for $n \in \{6\text{K}, 7\text{K}, \ldots, 11\text{K}\}$ filtered samples (out of 12K) for ADM diffusion model [Dhariwal and Nichol, 2021]. Our generative uncertainty performs on par with realism [Kynkäänniemi et al., 2019] and rarity scores [Han et al., 2023] in terms of FID (*left*), while exhibiting a weaker precision-recall trade-off (*middle* and *right*). We report mean values along with standard deviations over 5 runs with different random seeds.

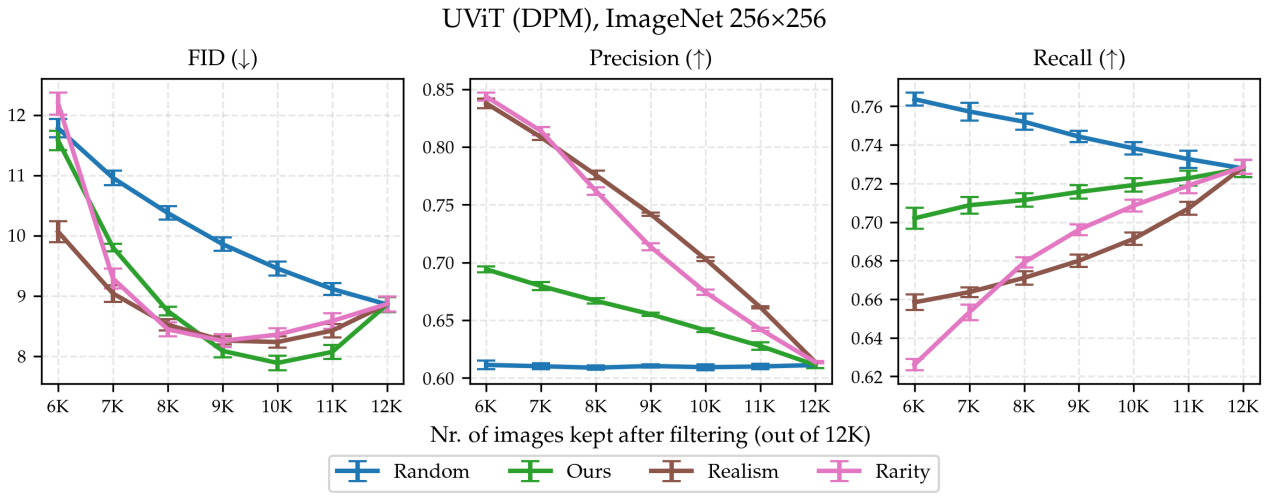

Figure 14: Image generation results for $n \in \{6\text{K}, 7\text{K}, \ldots, 11\text{K}\}$ filtered samples (out of 12K) for UViT diffusion model [Bao et al., 2023]. Our generative uncertainty performs on par with realism [Kynkäänniemi et al., 2019] and rarity scores [Han et al., 2023] in terms of FID (*left*), while exhibiting a weaker precision-recall trade-off (*middle* and *right*). We report mean values along with standard deviations over 5 runs with different random seeds.

To better understand the relationship between our generative uncertainty and non-uncertainty-based approaches such as realism [Kynkäänniemi et al., 2019] and rarity [Han et al., 2023] scores, we compute the Spearman correlation coefficient between different sample-level metrics on a set of 12K generated images from the experiment in Section 4.2. As shown in Figure 15, realism and rarity scores exhibit a strong correlation (< -0.8). This is unsurprising, as both scores are derived from the distance of a generated sample to a data manifold obtained using a reference dataset (e.g., a subset of training data or a separate validation dataset).[3]

In contrast, our generative uncertainty exhibits a weaker correlation (< 0.4) with both realism and rarity scores. We attribute this to the fact that our uncertainty primarily reflects the limited training data used in training diffusion models

[3]Such distance-based approaches are also commonly used to estimate prediction's quality in predictive models; see, for example, Van Amersfoort et al. [2020].

(i.e., epistemic uncertainty), rather than the distance to a reference dataset, as is the case for realism and rarity scores.

Next, we investigate whether combining different scores can improve the detection of low-quality generations. When combining two scores, we first rank the 12K images based on each score individually, then compute the combined ranking by summing the two rankings and re-ranking accordingly. The results, shown in Table 1, indicate that combining realism and rarity leads to minor or no improvements in FID (9.81 compared to 9.76 for realism alone on ADM. However, combining our generative uncertainty with either realism or rarity achieves the best FID performance (9.54 on ADM). These results suggest that ensembling scores that capture different aspects of generated sample quality is a promising direction for future research.

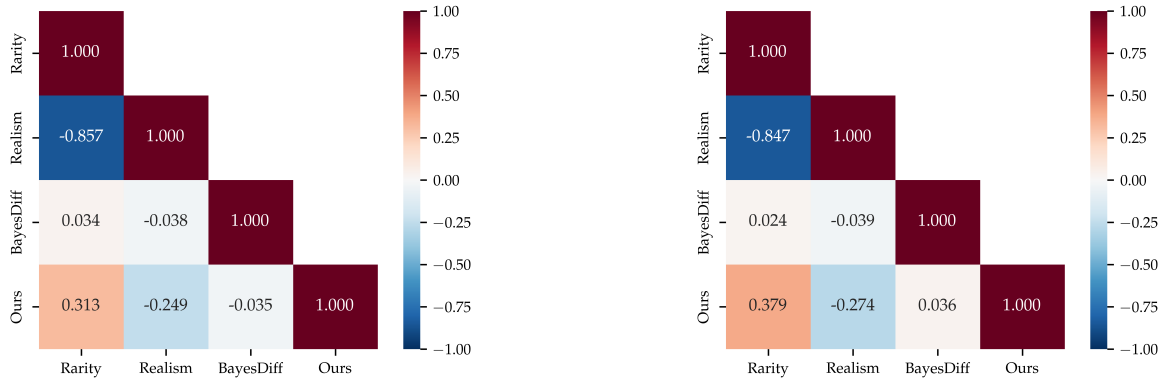

Figure 15: Spearman correlation coefficient between different sample quality metrics for 12K ImageNet images generated using ADM (*left*) and UViT (*right*).

Table 1: Image generation results for $n = 10$K filtered samples (out of 12K) based on combined metrics. Combining our generative uncertainty outperforms combining realism and recall in terms of FID. We report mean values along with standard deviation over 5 runs with different random seeds.

	ADM (DDIM), ImageNet 128×128			UViT (DPM), ImageNet 256×256		
	FID (↓)	**Precision** (↑)	**Recall** (↑)	**FID** (↓)	**Precision** (↑)	**Recall** (↑)
Realism + Rarity	9.81 ± 0.06	67.06 ± 0.29	66.73 ± 0.37	8.26 ± 0.07	69.01 ± 0.33	69.86 ± 0.36
Ours+ Realism	9.54 ± 0.04	66.41 ± 0.15	67.04 ± 0.47	7.60 ± 0.10	68.33 ± 0.09	69.75 ± 0.42
Ours + Rarity	9.56 ± 0.06	65.44 ± 0.26	67.36 ± 0.54	7.56 ± 0.12	67.48 ± 0.18	70.18 ± 0.40

C.6 CLASS-AVERAGED GENERATIVE UNCERTAINTY

To better understand the drop in sample diversity (recall) when using our generative uncertainty to filter low-quality samples in Figures 3&7, we analyze the distribution of average entropy per conditioning class. Specifically, for each of the 12K generated images, we randomly sample a conditioning class to mimic unconditional generation. As a result, all 1,000 ImageNet classes are represented among the 12K generated samples. Next, we compute our generative uncertainty for each sample and then average the uncertainties within each class. A plot of class-averaged uncertainties is shown in Figure 16. Since class-averaged uncertainties exhibit considerable variance, the class distribution in the 10K filtered samples deviates somewhat from that of the original 12K images, thereby explaining the reduction in diversity (recall).

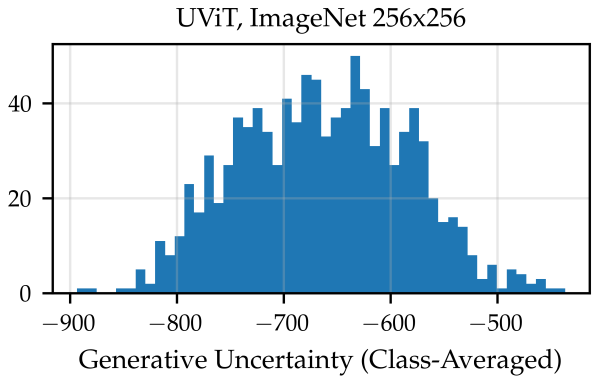

Figure 16: A histogram of class-averaged generative uncertainties for 12K generated samples using UViT.

While our primary focus in this work is on providing per-sample uncertainty estimates $u(\boldsymbol{z})$, we can also obtain uncertainty estimates for the conditioning variable $u(\boldsymbol{y})$ (e.g., a class label), by averaging over all samples corresponding to a particular $\boldsymbol{y} \in \mathcal{Y}$ as done in Figure 16. These estimates resemble the epistemic uncertainty scores proposed in DECU [Berry et al., 2024] and could be used to identify conditioning variables for which generated samples are likely to be of poor quality. We leave further exploration of generative uncertainty at the level of conditioning variables for future work.

C.7 FLOW MATCHING

To demonstrate that our generative uncertainty framework (Section 3) extends beyond diffusion models, we apply it here to the recently popularized flow matching approach [Lipman et al., 2023, Liu et al., 2023, Albergo et al., 2023]. Specifically, we consider a latent flow matching formulation [Dao et al., 2023] with a DiT backbone [Peebles and Xie, 2023]. For sampling, we employ a fifth-order Runge-Kutta ODE solver (`dopri5`). In Figure 17, we illustrate the samples with the highest and lowest generative uncertainty among 12K generated samples. On a filtered set of 10K images, our generative uncertainty framework achieves an FID of 10.48 and a precision of 64.71, significantly outperforming a random baseline, which yields an FID of 11.80 and a precision of 61.04.

Figure 17: Images with the highest (*left*) and the lowest (*right*) generative uncertainty among 12K generations using a latent flow matching model [Dao et al., 2023]. Uncertainty correlates with visual quality, as high-uncertainty samples exhibit numerous artefacts, whereas low-uncertainty samples resemble canonical images of their respective conditioning class.

D IMPLEMENTATION DETAILS

All our experiments can be conducted on a single $A100$ GPU, including the fitting of the Laplace posterior (Section 3.2). Our code is publicly available at `https://github.com/metodj/DIFF-UQ`.

Laplace Approximation When fitting a last-layer Laplace approximation (Section 3.2), we closely follow the implementation from BayesDiff [Kou et al., 2024]. Specifically, we use the empirical Fisher approximation with a diagonal factorization for Hessian computation. As the prior, we adopt a simple isotropic Gaussian distribution, $p(\theta) = \mathcal{N}(0, \gamma^{-1}I)$. The prior precision parameter and observation noise are fixed at $\gamma = 1$ and $\sigma = 1$, respectively. We report ablations for both parameters in Tables 3 and 4, finding that neither has a significant impact on the results. For Hessian computation, we utilize 1% of the training data for ImageNet 128×128 and 2% for ImageNet 256×256. Further details about the last layer of each diffusion model are provided in Table 2, where we observe that fewer than 1% of the parameters receive a 'Bayesian treatment'. We utilize `laplace`[4] library in our implementation.

	All Params.	**LL Params.**	**LL Name**
ADM	$\sim 421 \times 10^6$	$\sim 14 \times 10^3$	`out.2`
UViT	$\sim 500 \times 10^6$	$\sim 18 \times 10^3$	`decoder_pred`
DiT	$\sim 131 \times 10^6$	$\sim 1.2 \times 10^6$	`final_layer`

Table 2: Details of our last-layer (LL) Laplace approximation. The first column presents the total number of model parameters, while the second and third columns indicate the number of parameters in the last layer and its name, respectively

As discussed in Section 6, improving the quality of the Laplace approximation—such as incorporating both first and last layers instead of only the last layer [Daxberger et al., 2021b, Sharma et al., 2023] or optimizing Laplace hyperparameters (e.g., prior precision and observation noise) [Immer et al., 2021]—could further enhance the quality of generative uncertainty and represents a promising direction for future work.

Sampling with Generative Uncertainty For our main experiment in Section 4.2, we generate 12K images using the pretrained ADM model [Dhariwal and Nichol, 2021] for ImageNet 128×128 and the UViT model [Bao et al., 2023] for ImageNet 256×256. Following BayesDiff [Kou et al., 2024], we use a DDIM sampler [Song et al., 2021a] for the ADM model and a DPM-2 sampler [Lu et al., 2022] for the UViT model, both with $T = 50$ sampling steps.

To compute generative uncertainty (Algorithm 1), we first sample $M = 5$ sets of weights from the posterior $q(\theta|\mathcal{D})$. Then, for each of the initial 12K random seeds, we generate M additional samples. The same set of model weights $\{\theta_m\}_{m=1}^M$ is used for all 12K samples for efficiency reasons. For semantic likelihood (Eq. 7), we use a pretrained CLIP encoder [Radford et al., 2021] and set the semantic noise to $\sigma^2 = 0.001$.

Baselines For all baselines, we use the original implementation provided by the respective papers, except for [De Vita and Belagiannis, 2025], which we reimplemented ourselves since we were unable to get their code to run. Moreover, we use the default settings (e.g., hyperparameters) recommended by the authors for all baselines. For realism [Kynkäänniemi et al., 2019] and rarity [Han et al., 2023] we use InceptionNet [Szegedy et al., 2016] as a feature extractor and a subset of 50K ImageNet training images as the reference dataset. For samples where the rarity score is undefined (i.e., those that lie outside the estimated data manifold), we set it to `inf`.

Table 3: Ablation on prior precision parameter γ. Results for ADM model on ImageNet 128x128 dataset based on $n = 10$K filtered images (out of 12K).

γ	**FID** (↓)	**Precision** (↑)	**Recall** (↑)
1.0	10.04 ± 0.14	61.28 ± 0.23	69.55 ± 0.49
0.01	10.04 ± 0.12	61.14 ± 0.25	69.59 ± 0.52
0.1	10.05 ± 0.09	61.19 ± 0.21	69.75 ± 0.45
10.	10.01 ± 0.15	60.95 ± 0.28	69.71 ± 0.54
100.	10.06 ± 0.11	61.12 ± 0.26	69.62 ± 0.50

Table 4: Ablation on likelihood noise σ^2 parameter. Results for ADM model on ImageNet 128x128 dataset based on $n = 10$K filtered images (out of 12K).

σ^2	**FID** (↓)	**Precision** (↑)	**Recall** (↑)
0.1	10.30 ± 0.11	60.62 ± 0.31	69.99 ± 0.43
0.01	10.18 ± 0.04	61.18 ± 0.38	69.61 ± 0.59
0.001	10.04 ± 0.14	61.28 ± 0.23	69.55 ± 0.49
0.0001	10.01 ± 0.14	61.34 ± 0.23	69.50 ± 0.39
0.00001	10.10 ± 0.06	61.40 ± 0.28	69.53 ± 0.57

[4] `https://github.com/aleximmer/Laplace`

ADDITIONAL ACKNOWLEDGEMENTS AND DISCLAIMERS

Parts of this research were supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/ Interior Business Center (DOI/IBC) contract number 140D0423C0075. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/IBC, or the U.S. Government.