GENERATIVE UNCERTAINTY IN DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion and flow matching 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 the *generative uncertainty* of synthetic samples. We outline how to make Bayesian inference practical for large, modern generative models and introduce a new semantic likelihood 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., 2020a;b; Lipman et al., 2022) have recently pushed the boundaries of generative modeling due to their strong theoretical underpinnings and easy-to-scale nature. 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 impressive progress, state-of-the-art models can still generate low-quality images that contain artifacts or fail to align with the provided conditioning information. As a result, users deploying these models may need multiple generations to manually find a high-quality sample. This raises a key question: *how can we detect poor generations?*

Bayesian inference has long been applied to detect poor-quality predictions in predictive models (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. One notable exception is BayesDiff (Kou et al., 2024), which uses Bayesian uncertainty to filter out low-quality samples in diffusion models for natural images. However, BayesDiff's uncertainty estimates are not particularly indicative of a sample's actual quality and instead serve primarily as a detector of images with 'cluttered' backgrounds (Fig. 5).

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 (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 for filtering out poor samples (Kou et al., 2024; De Vita & Belagiannis, 2025). Additionally, we achieve competitive performance with non-uncertainty-based methods, such as realism scores (Kynkäänniemi et al., 2019) and rarity scores (Han et al., 2023) (Section 4.1).
3. We propose effective strategies to reduce the sampling overhead of Bayesian uncertainty (Section 4.2) and demonstrate the applicability of our framework beyond diffusion models by applying it to a (latent) flow matching model (Section D.7).

2 Background

2.1 Generative Modeling

Sampling in Generative Models Modern deep generative models like variational autoencoders (VAEs), generative adversarial networks (GANs), 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 $\theta \in \mathbb{R}^P$.

Diffusion Models The primary focus of this work is on diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2020b).[1] 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

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

where $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$ and $\{\beta_s\}_{s=1}^{T}$ 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. (2020b) or Karras et al. (2022)):

$$\mathcal{L}(\theta; \mathcal{D}) = \mathbb{E}_{t, \boldsymbol{x}_0, \epsilon}\left[\left|\left| f_\theta(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t) - \epsilon \right|\right|_2^2\right] \tag{1}$$

where $\mathcal{D} = \{\boldsymbol{x}_n\}_{n=1}^{N}$ denotes a training dataset of images. 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., 2020a).

2.2 Bayesian Deep Learning

Bayesian neural networks (BNNs) go beyond point predictions and allow for principled uncertainty quantification (MacKay, 1995; Kendall & 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 $\mathcal{D} = \{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{i=1}^{N}$ denote training data. A point-prediction model would find a single fixed set of parameters, $\hat{\psi} = \arg\max \mathcal{L}(\psi; \mathcal{D})$, that maximizes an objective function, $\mathcal{L}$. Meanwhile, a Bayesian Neural Network (BNN) specifies a prior, $p(\psi)$, over model parameters and defines a likelihood, $p(\boldsymbol{y}|h_\psi(\boldsymbol{x}))$, 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 (Murphy, 2022):

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

[1] Our framework extends beyond diffusion models and can be applied to other generative model families. See Figure TODO for a demonstration on flow matching models (Lipman et al., 2022).

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), 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:

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

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 like out-of-distribution (OOD) detection (Daxberger et al., 2021a) and active learning (Gal et al., 2017).

3 Generative Uncertainty via Bayesian Inference

While Bayesian neural networks (BNNs) have traditionally been applied to predictive models to estimate *predictive uncertainty*, we demonstrate how to apply them to diffusion to estimate *generative uncertainty* in this section (see Figure 4 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.

3.1 Bayesian Diffusion

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:

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

Here, we use $\boldsymbol{z}$ (with a slight abuse of notation) to denote the entire randomness involved in the diffusion sampling process.[2] Generative uncertainty is then defined as the variability of the posterior predictive:

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

where $H(\cdot)$ denotes the entropy. We propose a tractable estimator of entropy later in Eq. 8. We choose entropy as the variability measure 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 & Meger, 2023), can also be employed.

In the same way that the predictive uncertainty $u(\boldsymbol{x}_*)$, of a predictive model h_ψ provides insight into the quality of its prediction for a new test point $\boldsymbol{x}_*$, the generative uncertainty $u(\boldsymbol{z})$ of a diffusion 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 a significant concern. To address this, we adopt the Laplace approximation (MacKay,

[2]For example, in DDIM (Song et al., 2020a) and ODE sampling (Song et al., 2020b), $\boldsymbol{z} = \epsilon_T$, whereas in DDPM (Ho et al., 2020) and SDE sampling (Song et al., 2020b), $\boldsymbol{z} = \{\epsilon_T, \ldots, \epsilon_1\}$ with $\epsilon_t \sim \mathcal{N}(0, I)$.

1992; Shun & McCullagh, 1995) to approximate the 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:

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

where $\hat{\theta}$ represents the parameters of a pre-trained diffusion model, and Σ is the inverse Hessian of the diffusion (regression) 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.

We note 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 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.

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:

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

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, 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

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

where $\mathbf{e}(\mathbf{x}) \in \mathcal{S}$ is the vector of extracted semantic features.

By combining the (last-layer) Laplace approximate posterior and the semantic likelihood, we can now approximate the posterior predictive (Eq. 3) as

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

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

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).

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).

3.4 Epistemic Uncertainty

Bayesian uncertainty is commonly decomposed into two components: *aleatoric* and *epistemic* uncertainty (Hüllermeier & Waegeman, 2021; Smith et al., 2024). Aleatoric uncertainty represents the irreducible uncertainty inherent in the data-generating process, while epistemic uncertainty arises from observing only a limited amount of training data. In our framework, we fix the observation noise in the semantic likelihood (Eq. 7) to a small constant value (e.g., $\sigma = 0.001$). As a result, the generative uncertainty we capture is primarily epistemic in nature, reflecting uncertainty about the model parameters θ due to limited training data via $q(\theta|\mathcal{D})$. Extending our framework to additionally capture the aleatoric uncertainty of a generative process presents an interesting avenue for future research. Furthermore, since the parameters ϕ of the semantic feature extractor c_ϕ are kept fixed in the semantic likelihood, the resulting generative uncertainty $u(\boldsymbol{z})$ continues to reflect the epistemic uncertainty of the diffusion model parameters θ.

4 Experiments

In our experiments, we demonstrate that generative uncertainty is an effective method for detecting poor samples in diffusion models (Section 4.1). We also discuss the sampling overhead introduced by our Bayesian approach and show that it can be effectively minimized (Section 4.2). Finally, we extend our Bayesian framework beyond diffusion by applying it to detect low-quality samples in a (latent) flow matching model (Appendix D.7). Code to replicate all our experiments and figures is publicly available at `GITHUB REPO`.

4.1 Detecting Poor Generations

Table 1: Image generation results for 10K filtered samples (out of 12K) based on various metrics. Our generative uncertainty outperforms previously proposed uncertainty-based approaches in terms of image quality (AU (De Vita & Belagiannis, 2025), BayesDiff (Kou et al., 2024)), as indicated by higher FID and precision scores, and is competitive with non-uncertainty methods (Realism (Kynkäänniemi et al., 2019), Rarity (Han et al., 2023)). We report mean values along with standard deviations over 3 runs with different random seeds.

	ADM (DDIM), ImageNet 128×128			UViT (DPM), ImageNet 256×256		
	FID (↓)	**Precision** (↑)	**Recall** (↑)	**FID** (↓)	**Precision** (↑)	**Recall** (↑)
Random	11.31 ± 0.07	58.90 ± 0.36	70.68 ± 0.38	9.46 ± 0.12	60.94 ± 0.24	73.82 ± 0.33
BayesDiff	11.20 ± 0.05	58.80 ± 0.05	70.62 ± 0.32	9.16 ± 0.17	61.77 ± 0.19	73.72 ± 0.38
AU	11.39 ± 0.05	58.82 ± 0.42	70.70 ± 0.38	9.20 ± 0.12	61.80 ± 0.33	73.46 ± 0.24
Ours	10.14 ± 0.08	61.26 ± 0.26	69.60 ± 0.49	7.89 ± 0.12	64.14 ± 0.17	71.92 ± 0.35
Realism	9.76 ± 0.04	67.95 ± 0.19	66.32 ± 0.40	8.24 ± 0.09	70.29 ± 0.15	69.12 ± 0.32
Rarity	10.09 ± 0.02	64.99 ± 0.16	67.73 ± 0.47	8.37 ± 0.11	67.21 ± 0.10	67.76 ± 0.48

To evaluate whether our newly introduced generative uncertainty can be used to detect low-quality generations, we follow the experimental setup from prior work on uncertainty-based filtering (Kou et al., 2024; De Vita & Belagiannis, 2025). Specifically, we generate 12K samples using a given diffusion model and compute the uncertainty estimate for each sample. We then select the 10K 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 10K 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

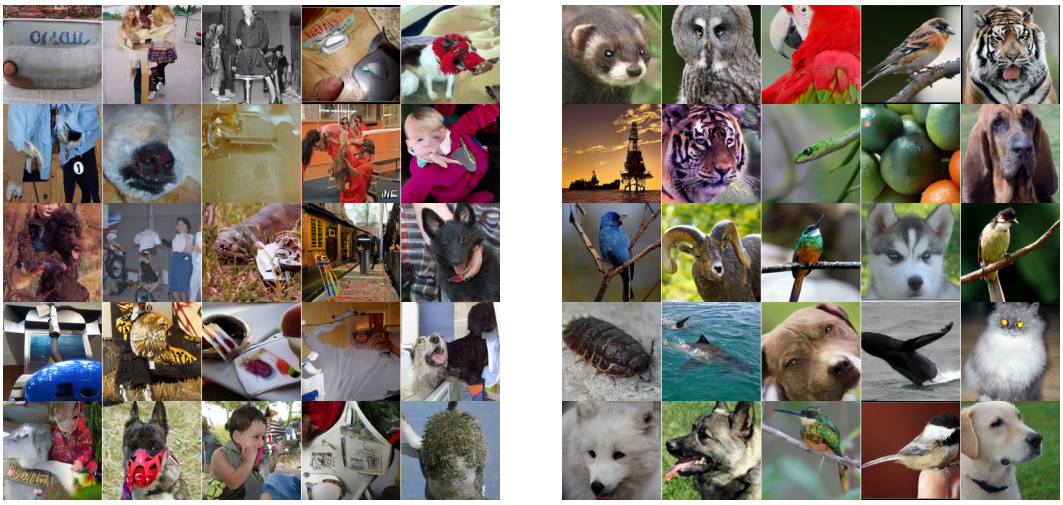

Figure 1: Images with the highest (*left*) and the lowest (*right*) generative uncertainty (Eq. 8) among 12K generations using a UViT diffusion model (Bao et al., 2023). Generative uncertainty correlates with visual quality, as high-uncertainty samples exhibit numerous artefacts, whereas low-uncertainty samples resemble canonical images of their respective conditioning class.

Figure 2: Images with the highest (*bottom*) and the lowest (*top*) generative uncertainty among 128 generations using a UViT diffusion model (Bao et al., 2023) for 2 classes: `black swan` (*left*) and `Tibetan terrier` (*right*).

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). Additional implementation details are provided in Appendix E.

Baselines We first compare our proposed generative uncertainty to existing uncertainty-based approaches for detecting low-quality samples: BayesDiff (Kou et al., 2024) and the aleatoric uncertainty (AU) approach proposed by De Vita & Belagiannis (2025). BayesDiff estimates epistemic uncertainty in diffusion models using a last-layer Laplace approximation and tracks this uncertainty throughout the entire sampling process via their variance propagation method. In contrast, De Vita & Belagiannis (2025) computes aleatoric uncertainty by measuring the sensitivity of intermediate diffusion scores to random perturbations. Unlike our approach, both methods estimate uncertainty directly in pixel space.

Importantly, we also compare our method against non-uncertainty-based sample-level metrics, 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 & Belagiannis, 2025) has not considered such comparisons, which we believe are essential for assessing the practical utility of uncertainty-based filtering.

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 Table 1. 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 (Bao et al., 2023), a subset of 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 1. 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 Table 1, 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 5 we also depict the 25 "worst" and "best" samples according to the uncertainty estimate from BayesDiff (Kou et al., 2024). 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, in Figure 7 we perform an ablation where we compute our generative uncertainty directly in the pixel-space. It is clear that without semantic likelihood, our uncertainty becomes overly sensitive to the background pixels in the same way as in BayesDiff.

Returning to Table 1, we observe that filtering based on our generative uncertainty results in some loss of sample diversity, as evidenced by lower recall scores (e.g., 73.82 for Random vs. 71.92 for our method on the UViT model). We attribute this to the fact that, in our main experiment, 12K images are generated unconditionally.[3] As a result, all 1000 ImageNet classes are represented. Since certain classes produce images with higher uncertainty (see Appendix D.6 for a detailed analysis), filtering based on uncertainty inevitably alters the class distribution among the selected samples. We expect this issue to be less pronounced in conditional generation (see Figure 2). 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 & Salimans, 2022).

Lastly, we compare our proposed method with non-uncertainty-based approaches—a comparison missing in prior literature (Kou et al., 2024; De Vita & Belagiannis, 2025). For realism (Kynkäänniemi et al., 2019), we retain the 10K images with the highest scores, whereas for rarity (Han et al., 2023), we keep those with the lowest scores. As shown in Table 1, our generative uncertainty is the only uncertainty-based method that approaches realism and rarity in terms of FID (e.g., 7.89 for ours vs. 8.24 for realism and 8.37 for rarity on UViT). However, a large gap remains in precision (e.g., 64.14 for ours vs. 70.29 for realism and 67.21 for rarity on UViT). Notably, realism and rarity sacrifice the most sample diversity, as indicated by their lowest recall scores (e.g., 69.12 for realism and 67.76 for rarity on UViT).

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

Furthermore, Table 2 shows that our score can be effectively combined with realism or rarity scores. Specifically, 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 10.

4.2 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 Table 1, 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.

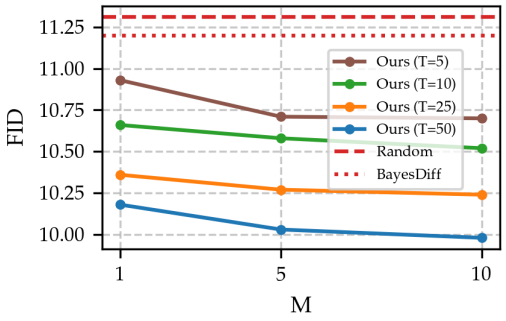

Figure 3: FID results for 10K ImageNet-filtered images using our generative uncertainty on ADM model (Dhariwal & Nichol, 2021). We vary the number of Monte Carlo samples M and diffusion sampling steps T (see Algorithm 1). By default, we use $M = 5$ with $T = 50$, incurring an additional 250 NFEs for uncertainty estimation. Encouragingly, setting $M = 1$ and $T = 25$ still achieves competitive performance while reducing the sampling overhead by 10x.

The most straightforward approach is to reduce the number of Monte Carlo samples M. Encouragingly, reducing M to as low as 1 still achieves highly competitive performance (see Figure 3). 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. Concretely, using the ADM model (Dhariwal & 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 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 optimizing diffusion hyperparameters via marginal likelihood using the Laplace approximation.

REFERENCES

Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint 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 *International Conference on Machine Learning*, pp. 717–730. PMLR, 2022.

Javier Antorán, James Allingham, and José Miguel Hernández-Lobato. Depth uncertainty in neural networks. *Advances in neural information processing systems*, 33:10620–10634, 2020.

Julyan Arbel, Konstantinos Pitas, Mariia Vladimirova, and Vincent Fortuin. A primer on bayesian neural networks: review and debates. *arXiv preprint 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22669–22679, 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 preprint arXiv:2308.13498*, 2023.

Lucas Berry, Axel Brando, and David Meger. Shedding light on large generative networks: Estimating epistemic uncertainty in diffusion models. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2024.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pp. 1613–1622. PMLR, 2015.

Matthew Albert Chan, Maria J Molina, and Christopher Metzler. Estimating epistemic and aleatoric uncertainty with a single model. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

François Cornet, Grigory Bartosh, Mikkel N Schmidt, and Christian A Naesseth. Equivariant neural diffusion for molecule generation. In *38th Conference on Neural Information Processing Systems*, 2024.

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

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

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux-effortless bayesian deep learning. *Advances in Neural Information Processing Systems*, 34:20089–20103, 2021a.

Erik Daxberger, Eric Nalisnick, James U Allingham, Javier Antorán, and José Miguel Hernández-Lobato. Bayesian deep learning via subnetwork inference. In *International Conference on Machine Learning*, pp. 2510–2521. PMLR, 2021b.

Michele De Vita and Vasileios Belagiannis. Diffusion model guided sampling with pixel-wise aleatoric uncertainty estimation. *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 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 *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=FPnUhsQJ5B.

Gianni Franchi, Dat Nguyen Trong, Nacim Belkhir, Guoxuan Xia, and Andrea Pilzer. Towards understanding and quantifying uncertainty for text-to-image generation, 2024. URL https://arxiv.org/abs/2412.03178.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International conference on machine learning*, pp. 1183–1192. PMLR, 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. *arXiv preprint arXiv:2210.08933*, 2022.

Shuyang Gu, Jianmin Bao, Dong Chen, and Fang Wen. Giqa: Generated image quality assessment. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pp. 369–385. Springer, 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 *The Eleventh International Conference on Learning Representations (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. *Advances in neural information processing systems*, 30, 2017.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Emiel Hoogeboom, Vıctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International conference on machine learning*, pp. 8867–8887. PMLR, 2022.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506, 2021.

Jaehui Hwang, Junghyuk Lee, and Jong-Seok Lee. Anomaly score: Evaluating generative models and individual generated images based on complexity and vulnerability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8754–8763, 2024.

Alexander Immer, Matthias Bauer, Vincent Fortuin, Gunnar Rätsch, and Khan Mohammad Emtiyaz. Scalable marginal likelihood estimation for model selection in deep learning. In *International Conference on Machine Learning*, pp. 4563–4573. PMLR, 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*, 17(2):29–48, 2022.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Siqi Kou, Lei Gan, Dequan Wang, Chongxuan Li, and Zhijie Deng. Bayesdiff: Estimating pixel-wise uncertainty in diffusion via bayesian inference. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Being bayesian, even just a bit, fixes overconfidence in relu networks. In *International conference on machine learning*, pp. 5436–5446. PMLR, 2020.

Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*, 2022.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 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 *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=XHTl2k1LYk.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

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. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022.

Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *Advances in Neural Information Processing Systems*, 36, 2024.

David JC MacKay. Bayesian interpolation. *Neural computation*, 4(3):415–447, 1992.

David JC MacKay. Bayesian neural networks and density networks. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 354(1):73–80, 1995.

Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. *Advances in neural information processing systems*, 32, 2019.

D Mandt, Matthew D Hoffman, and David M Blei. Stochastic gradient descent as approximate bayesian inference. *Journal of Machine Learning Research*, 18(134):1–35, 2017.

Kevin P Murphy. *Probabilistic machine learning: an introduction*. MIT press, 2022.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? *arXiv preprint arXiv:1810.09136*, 2018.

Koichi Namekata, Amirmojtaba Sabour, Sanja Fidler, and Seung Wook Kim. Emerdiff: Emerging pixel-level semantic knowledge in diffusion models. *arXiv preprint arXiv:2401.11739*, 2024.

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

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 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 *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Yunus Saatci and Andrew G Wilson. Bayesian gan. *Advances in neural information processing systems*, 30, 2017.

Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. *Advances in neural information processing systems*, 31, 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 *International Conference on Artificial Intelligence and Statistics*, pp. 7694–7722. PMLR, 2023.

Zhenming Shun and Peter McCullagh. Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 57(4):749–760, 1995.

Freddie Bickford Smith, Jannik Kossen, Eleanor Trollope, Mark van der Wilk, Adam Foster, and Tom Rainforth. Rethinking aleatoric and epistemic uncertainty. *arXiv preprint arXiv:2412.20892*, 2024.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

C Szegedy. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 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 *International Conference on Machine Learning*, pp. 33940–33960. PMLR, 2023.

Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models, 2016. URL `https://arxiv.org/abs/1511.01844`.

Jakub M. Tomczak. *Deep Generative Modeling*. Springer, Germany, February 2022. ISBN 978-3-030-93157-5. doi: 10.1007/978-3-030-93158-2.

Ba-Hien Tran, Babak Shahbaba, Stephan Mandt, and Maurizio Filippone. Fully bayesian autoencoders with latent sparse gaussian processes. In *International Conference on Machine Learning*, pp. 34409–34430. PMLR, 2023.

Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *International conference on machine learning*, pp. 9690–9700. PMLR, 2020.

Andrew Gordon Wilson. The case for bayesian deep learning. *arXiv preprint 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*, 10:e1905, 2024.

Cheng Zhang, Judith Bütepage, Hedvig Kjellström, and Stephan Mandt. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):2008–2026, 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? *arXiv preprint arXiv:2406.03070*, 2024.

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 *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 450–459, 2024.

APPENDIX

The supplementary material is organized as follows:

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

A GENERATIVE UNCERTAINTY: METHOD AND ALGORITHM

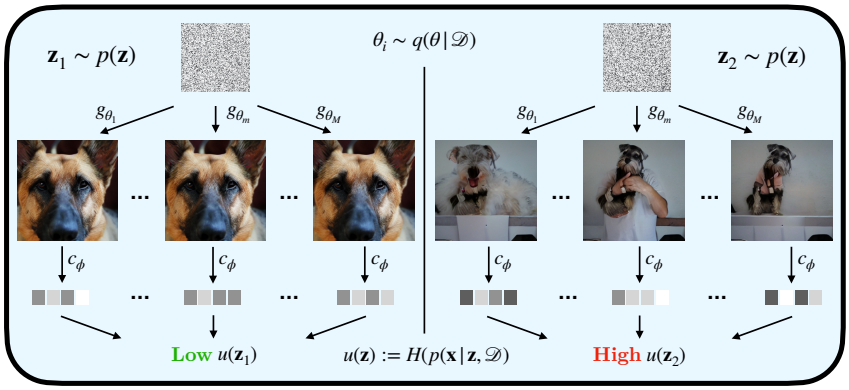

Figure 4: Demonstration of how we compute *generative uncertainty* for each random noise $\boldsymbol{z}$. We sample M sets of model parameters from our posterior distribution $q(\theta|\mathcal{D})$ and generate M images. Then, we evaluate the semantic likelihood for each by computing feature embeddings with a pretrained encoder c_ϕ (e.g., CLIP) and take the uncertainty (e.g. entropy) over these embeddings. Random noises $\boldsymbol{z}$ with low uncertainty (*left*) tend to lead to consistent, high quality generations while random noises with high uncertainty (*right*) lead to poor, discordant generations.

Algorithm 1: Diffusion Sampling with Generative Uncertainty

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. 7)
Compute the entropy $u(\boldsymbol{z}) = H(p(\boldsymbol{x}|\boldsymbol{z}, \mathcal{D}))$
return $\hat{\boldsymbol{x}}_0, u(\boldsymbol{z})$

B 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 & Belagiannis (2025), which captures aleatoric uncertainty via the sensitivity of diffusion score estimates. Our work extends both by proposing a more general (applicable beyond diffusion), simpler approach (requiring no sampling modifications), and a more effective (see Section 4.1) uncertainty framework.

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 operates *post-hoc* with any pretrained diffusion model via the Laplace approximation (Daxberger et al., 2021a). A recent approach, PUNC (Franchi et al., 2024), focuses specifically 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 & Wilson (2017) where a Bayesian version of a GAN is proposed, showing improvements for semi-supervised learning, and Daxberger & 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., 2018). Since diffusion models can be interpreted as neural ODEs (Song et al., 2020b), 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., 2024) 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.

C 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 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., 2022; Yi et al., 2024). Exploring whether insights from the literature on uncovering semantic features in diffusion models (Kwon et al., 2022; 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.

D Additional Results

D.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 5. These samples are drawn from the same set of 12K ImageNet "unconditional" images generated

using the UViT model (Bao et al., 2023) as in Figure 1. 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 Table 1), 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 5 looks quite good despite being assigned (very) high uncertainty.

Similarly, in Figure 6, we show low- and high-uncertainty samples according to BayesDiff for the same set of 128 images per class as in Figure 2. Once again, we observe that BayesDiff's uncertainty metric is less informative regarding a sample's visual quality compared to our generative uncertainty.

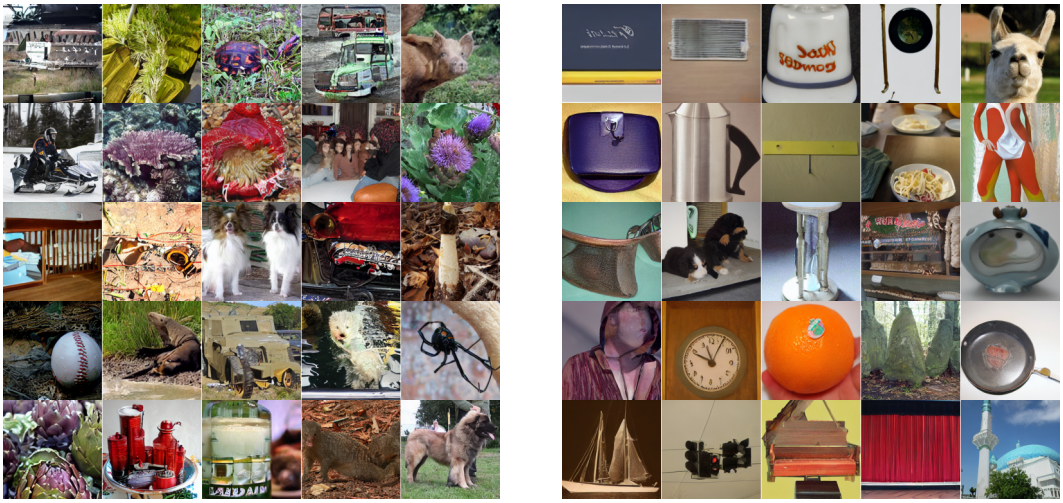

Figure 5: Images with the highest (*left*) and the lowest (*right*) BayesDiff uncertainty (Kou et al., 2024) among 12K generations using a UViT diffusion model (Bao et al., 2023). BayesDiff uncertainty correlates poorly with visual quality and is overly sensitive to the background pixels. Same set of 12K generated images is used as in Figure 1 to ensure a fair comparison.

Figure 6: Images with the highest (*bottom*) and the lowest (*top*) BayesDiff uncertainty (Kou et al., 2024) among 128 generations using a UViT diffusion model (Bao et al., 2023) for 2 classes: `black swan` (*left*) and `Tibetan terrier` (*right*). Same set of 128 generated images per class is used as in Figure 2 to ensure a fair comparison.

D.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 7 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 (Kou et al., 2024) (see Appendix D.1). This highlights the necessity of using semantic likelihood to obtain uncertainty estimates that are truly informative about the visual quality of generated samples.

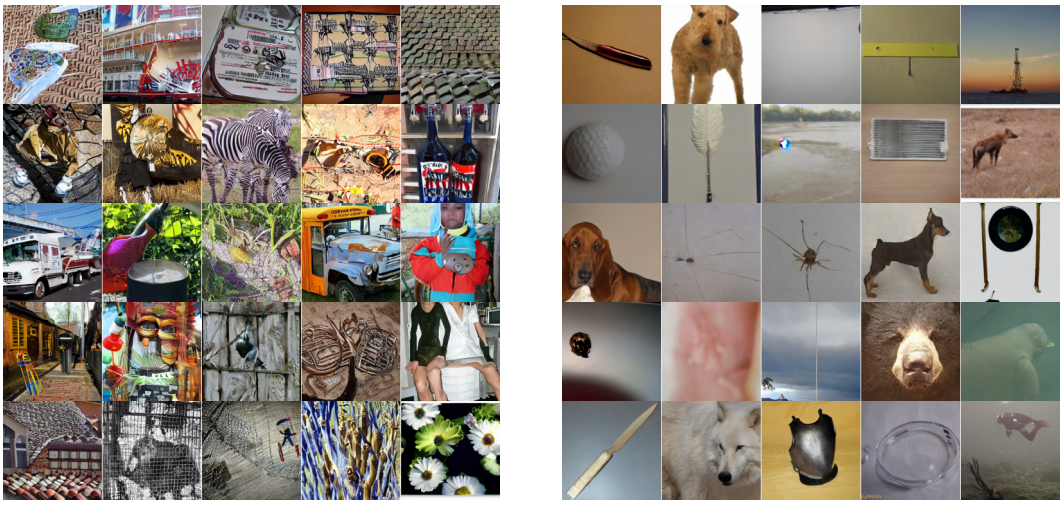

Figure 7: Images with the highest (*left*) and the lowest (*right*) 'pixel-space' generative uncertainty among 12K generations using a UViT diffusion model (Bao et al., 2023). Pixel-space uncertainty correlates poorly with visual quality and is overly sensitive to the background pixels. Same set of 12K generated images is used as in Figure 1 to ensure a fair comparison.

Figure 8: Pixel-wise uncertainty based on our generative uncertainty for 5 generated samples using UViT (Bao et al., 2023).

D.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 8 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 & 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 8). 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 (see Table 1)—enable more meaningful comparisons between different approaches.

D.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. In the same way as in our other comparisons, we retain the 10K generated images with highest likelihood. We utilize the

implementation in Dhariwal & 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 9. 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).

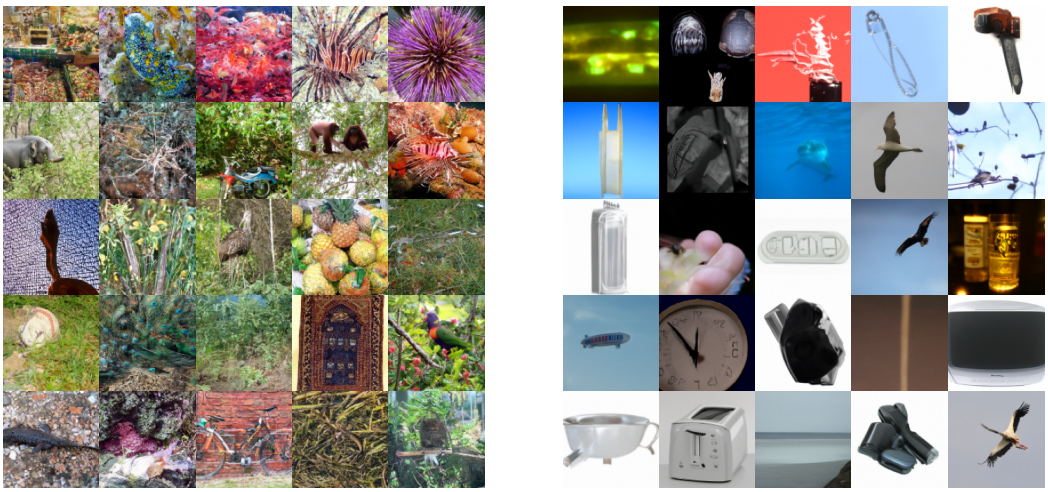

Figure 9: The 25 "Worst" (*left*) and "Best" (*right*) samples generated by ADM trained on ImageNet 128x128 selected by lowest and highest likelihood among 12K generations.

D.5 Comparison with Realism & Rarity

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 4.1. As shown in Figure 10, 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).[4]

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 (i.e., epistemic uncertainty, see Section 3.4), 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 2, 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 (Dhariwal & Nichol, 2021)). 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.

D.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 Table 1, we analyze the distribution of average entropy per conditioning class. Specifically, for each of the 12K generated images, we randomly sample a conditioning class

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

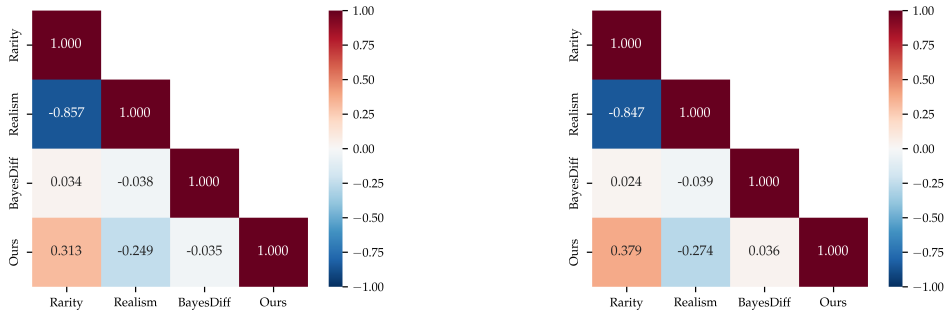

Figure 10: Spearman correlation coefficient between different sample quality metrics for 12K ImageNet images generated using ADM (Dhariwal & Nichol, 2021) (*left*) and UViT (Bao et al., 2023) (*right*).

Table 2: Image generation results for 10K 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 3 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

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 (i.e., entropy; see Eq. 8) for each sample and then average the uncertainties within each class. A plot of class-averaged uncertainties is shown in Figure 11. 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).

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 11. 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.

D.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., 2022; Liu et al., 2022; Albergo et al., 2023). Specifically, we consider a latent flow matching formulation (Dao et al., 2023) with a DiT backbone (Peebles & Xie, 2023). For sampling, we employ a fifth-order Runge-Kutta ODE solver (`dopri5`). In Figure 12, 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

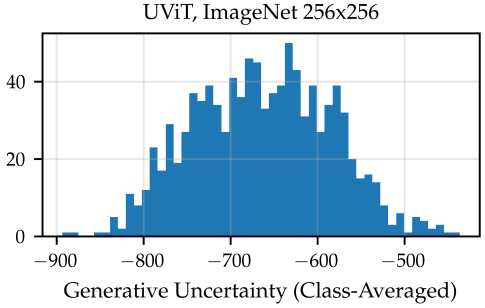

Figure 11: A histogram of class-averaged generative uncertainties for 12K generated samples using UViT (Bao et al., 2023).

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 12: Images with the highest (*left*) and the lowest (*right*) generative uncertainty (Eq. 8) among 12K generations using a latent flow matching model (Dao et al., 2023). Generative uncertainty correlates with visual quality, as high-uncertainty samples exhibit numerous artefacts, whereas low-uncertainty samples resemble canonical images of their respective conditioning class.

E Implementation Details

All our experiments can be conducted on a single $A100$ GPU, including the fitting of the Laplace posterior (Section 3.2). Code for reproducing our experiments is publicly available at `GITHUB REPO`.

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. The prior precision parameter and observation noise are fixed at $\gamma = 1$ and $\sigma = 1$, respectively. 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 3, where we observe that fewer than 1% of the parameters receive a 'Bayesian treatment'. For code implementation, we rely on the `laplace`[5] library (Daxberger et al., 2021a).

	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 3: 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 C, 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 4.1, we generate 12K images using the pretrained ADM model (Dhariwal & Nichol, 2021) for ImageNet 128×128

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

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., 2020a) 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 & 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`.