UNCERTAINTY-AWARE DISTRIBUTION-TO-DISTRIBUTION FLOW MATCHING FOR SCIENTIFIC IMAGING

Anonymous authors
Paper under double-blind review

ABSTRACT

Distribution-to-distribution generative models support scientific imaging tasks ranging from modeling cellular perturbation responses to translating medical images across conditions. Trustworthy generation requires *reliability*, or generalization across labs, devices, and experimental conditions, and *accountability*, or detecting out-of-distribution cases where predictions may be unreliable. We leverage Stochastic Flow Matching (SFM), a marginal-preserving stochastic extension of flow matching for improved generalization under distribution shift. SFM augments deterministic flows with a diffusion term together with a learned score-based drift correction, retaining the learned transport marginals while modeling conditional variability. Building on this SFM framework, we introduce Bayesian Stochastic Flow Matching (BSFM) as a companion uncertainty quantification mechanism and develop AVUQ (Antithetic Variance-reduction Uncertainty Quantification) to approximately estimate epistemic and aleatoric uncertainty via sample-efficient antithetic sampling with approximate posterior inference. We further use AVUQ to yield anomaly scores for unreliable generation detection. Experiments on cellular imaging (BBBC021, JUMP) and brain fMRI (Theory of Mind) across diverse unseen scenarios show that SFM improves generalization while AVUQ provides effective uncertainty-based anomaly scores under practical sampling budgets.

1 INTRODUCTION

Distribution-to-distribution generative image models predict cellular morphology after chemical or genetic perturbations (Zhang et al., 2025b), map resting-state to task-evoked brain activity (Kwon et al., 2025; Kan et al., 2022), and translate medical images between control and treatment conditions (Li et al., 2023b; Arslan et al., 2025). By learning transformations between source and target distributions, they capture biological responses to interventions, neural activity shifts, and manifestations of pathology. Scientific deployment requires more than generative quality (Huang et al., 2025; Blau et al., 2024); it also requires **reliability** across unseen perturbation types, cell lines, assay plates, laboratories, imaging devices, and protocols; and **accountability**, the capacity to detect out-of-distribution (OOD) cases and unreliable generations. Otherwise, misleading predictions can prompt costly validation of false hypotheses, create patient safety concerns, or miss scientific opportunities. In drug discovery, confidently predicted responses to novel compounds could misdirect expensive screening campaigns. Fig. 1 (left) illustrates these challenges.

Flow matching (Lipman et al., 2022; Zhang et al., 2025b) is a natural backbone for distribution-to-distribution generation because it learns continuous transformations via neural ordinary differential equations (ODEs). However, standard flow matching is deterministic: for a fixed input and condition, it produces a single generated output. This can limit generalization under scientific distribution shifts that require the model to represent conditional variability rather than a single transport path. A naive stochastic extension that simply injects diffusion noise can corrupt the learned marginals and degrade generation quality. Thus, a key technical challenge is to add stochasticity in a way that improves robustness while preserving the learned distribution-to-distribution mapping.

We show that **Stochastic Flow Matching (SFM)** (Song et al., 2020; Karras et al., 2022; Su et al., 2025) can be used to address this challenge by adding stochasticity via a marginal-preserving

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

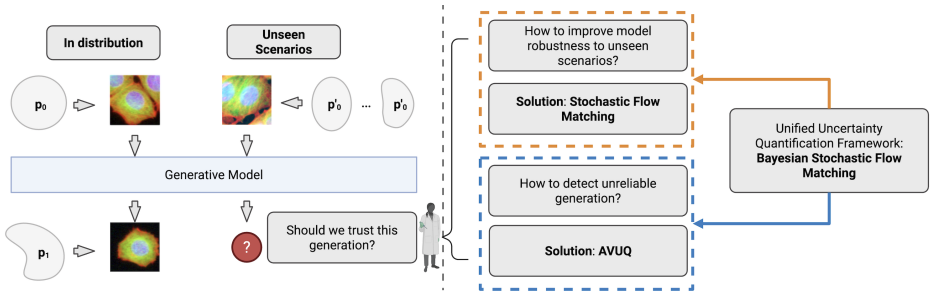


Figure 1: Trustworthy distribution-to-distribution generative modeling. **Left:** In-domain samples are drawn from the source distribution p_0, while unseen samples may arise from shifted source distributions $\{p'_0\}$. **Right:** Our approach centers on Stochastic Flow Matching (SFM) to improve model generalization, with Bayesian Stochastic Flow Matching (BSFM) and AVUQ providing approximate aleatoric and epistemic uncertainty estimates for unreliable generation detection.

stochastic extension of deterministic flow matching. SFM introduces a diffusion term together with a learnable score-based drift correction, so stochastic sampling can model conditional variability without changing the target marginals of the learned transport. Although this model has been introduced before for generative modeling, we are the first to show that SFM improves generalization to unseen scenarios in distribution-to-distribution scientific imaging: it retains the efficiency and regression-based training of flow matching while improving robustness to unseen scenarios.

Uncertainty quantification (UQ) provides a complementary accountability mechanism, but scaling UQ to high-dimensional image generators is computationally challenging. **Bayesian Stochastic Flow Matching** (BSFM) provides a promising UQ framework, but exact Bayesian inference is intractable for the velocity and score networks, and naive nested sampling requires many costly SDE solves to obtain stable estimates. We therefore develop **AVUQ** (Antithetic Variance-reduction Uncertainty Quantification) by leveraging antithetic sampling for reduced-variance, sample-efficient estimation of both aleatoric and epistemic uncertainty terms. Although AVUQ can be applied with any method of generating approximate posterior samples, in this paper we use MC-Dropout (MCD) as a scalable posterior approximation. We show empirically that this MCD instantiation outperforms other approximate inference methods. The key contribution, however, is the sample-efficient variance reduction provided by AVUQ; without it, nested sampling would be computationally prohibitive. The resulting approximate estimates of aleatoric and epistemic uncertainty serve as anomaly scores for unreliable generation detection without architectural changes or additional training.

Through extensive experiments on cellular imaging datasets (BBBC021, JUMP) and an fMRI dataset ('Theory of Mind', ToM) across diverse unseen scenarios—novel perturbations, different laboratory settings, novel cell lines, unseen assay plates, and low-performing ToM subjects—we demonstrate the effectiveness of SFM and AVUQ. Specifically, SFM improves reliability through enhanced generalization under distribution shifts, while AVUQ provides sample-efficient uncertainty estimates that support effective unreliable generation detection. Together, these components provide a practical recipe for more trustworthy distribution-to-distribution generative modeling in scientific applications (Fig. 1, right). In summary, our contributions are:

- We show **Stochastic Flow Matching (SFM)**, a marginal-preserving stochastic extension of flow matching can be used to improve generalization under distribution shifts.
- Proposing **AVUQ** (Antithetic Variance-reduction Uncertainty Quantification) in the **Bayesian Stochastic Flow Matching (BSFM)** framework, which provides approximate and sample-efficient estimates of aleatoric and epistemic uncertainty.
- Demonstration across scientific imaging benchmarks that SFM improves generation quality in unseen scenarios and that AVUQ provides effective uncertainty-based signals for unreliable generation detection in OOD scenarios.

2 Related Work

Distribution-to-distribution Generative Models. Distribution-to-distribution models map between two data distributions rather than from noise to data, with applications across scientific

domains (Zhang et al., 2025b; Kwon et al., 2025; Kan et al., 2022; Li et al., 2023b; Arslan et al., 2025). Approaches include flow-based models (Lipman et al., 2022; 2024), diffusion bridges (Shi et al., 2023; Zhou et al., 2023; De Bortoli et al., 2021), and generative adversarial networks (Zhu et al., 2017; Park et al., 2020; Kim et al., 2023). Flow matching trains by velocity-field regression and generates through ODE integration. Extensions use classifier-free guidance for conditional generation (Zhang et al., 2025b) and stochastic perturbations for improved in-distribution performance (Su et al., 2025). Despite high in-distribution fidelity, behavior under distribution shifts remains unexplored, motivating UQ for unreliable generation.

Uncertainty Quantification in Generative Models. Diffusion-model UQ has drawn growing interest. Post-hoc last-layer Laplace methods such as BayesDiff (Kou et al., 2023) and Jazbec et al. (Jazbec et al., 2025) quantify epistemic uncertainty; denoiser ensembles such as DECU (Berry et al., 2024) capture model variability. Posterior sampling quantifies aleatoric uncertainty (De Vita & Belagiannis, 2025). Applications include inverse imaging (Xie & Li, 2022; Feng et al., 2023; Wu et al., 2024) and text-to-image generation (Franchi et al., 2025); conformal methods (Teneggi et al., 2023; Ekmekci & Cetin, 2025) provide distribution-free coverage. Hyper-Diffusion (Chan et al., 2024) decomposes both uncertainty types in one hypernetwork but uses costly nested sampling. Existing methods primarily address *noise-to-data* diffusion generation; UQ for *flow-based distribution-to-distribution* generation remains unexplored.

Out-of-Distribution Detection in Generative Models. Generative-model OOD detection commonly uses reconstruction or manifold consistency, including diffusion inpainting (Liu et al., 2023), denoising trajectories (Graham et al., 2023a), semantic mismatch (Gao et al., 2023), and projection regret (Choi et al., 2023). Other cues include diffusion-path curvature (Heng et al., 2024), covariance spectra (Shoushtari et al., 2025), and norm-guided residuals (Zhang et al., 2025a). Applications span medical imaging (Graham et al., 2023b; Bercea et al., 2024; Linmans et al., 2024), road scenes (Galesso et al., 2024), zero-shot (Abdi et al., 2025), and trend-based (Kim et al., 2024) anomaly detection. These methods define OOD globally against one data distribution in noise-to-data or image-to-manifold settings. We address distribution-to-distribution settings, where OOD is conditional on the source distribution; this source-aware regime has not been addressed by diffusion or flow-matching approaches.

3 PRELIMINARIES

3.1 DISTRIBUTION-TO-DISTRIBUTION IMAGE GENERATION

Let p_0 be a source distribution and p_1 a target distribution over the image space $\mathcal{X}$. The goal of distribution-to-distribution image generation is to learn a generative model mapping p_0 to p_1. Given an image $\boldsymbol{x}_0 \sim p_0$, the model generates $\boldsymbol{x}_1 \sim p_1$ representing the transformed image in the target distribution. Optionally, we can incorporate a condition space $\mathcal{C}$ to enable conditional generation, where the target distribution becomes $p_1(\cdot|c)$ for $c \in \mathcal{C}$. We then learn a conditional generative model $p(\boldsymbol{x}_1|\boldsymbol{x}_0, c)$ that maps $\boldsymbol{x}_0 \sim p_0$ to $\boldsymbol{x}_1 \sim p_1(\cdot|c)$. In cellular imaging, p_0 and p_1 represent the distributions of unperturbed cell images and images under drug perturbation c, respectively. The model predicts morphological changes induced by the perturbation.

3.2 FLOW MATCHING-BASED GENERATIVE MODELING

Flow matching-based generative models learn invertible mappings between p_0 and p_1 via continuous transformations. Given pairs of samples from these distributions, flow matching learns a time-dependent velocity field using a neural network $\boldsymbol{v}_\theta : \mathcal{X} \times [0,1] \to \mathcal{X}$ that defines the instantaneous direction and magnitude of change at each point. The transformation process follows the ODE:

$$\frac{d\boldsymbol{x}_t}{dt} = \boldsymbol{v}_\theta(\boldsymbol{x}_t, t), \quad \boldsymbol{x}_0 \sim p_0, \boldsymbol{x}_1 \sim p_1, t \in [0,1]. \tag{1}$$

We employ the rectified flow formulation (Liu et al., 2022), which yields a straight-line path:

$$\boldsymbol{x}_t = (1-t)\boldsymbol{x}_0 + t\boldsymbol{x}_1, \quad t \sim \mathcal{U}[0,1]. \tag{2}$$

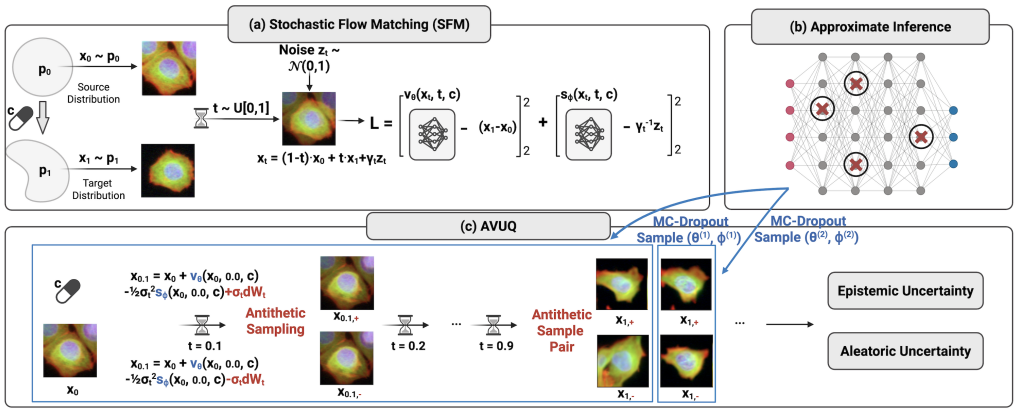


Figure 2: Core components of the Bayesian Stochastic Flow Matching (BSFM) framework: (a) Stochastic Flow Matching (SFM) training process. (b) MC-Dropout as one approximate inference method that can be used by AVUQ. (c) AVUQ with antithetic sampling.

The linear path has velocity field $\boldsymbol{v}(\boldsymbol{x}_t, t) = d\boldsymbol{x}_t/dt = \boldsymbol{x}_1 - \boldsymbol{x}_0$, representing the optimal transport direction at each point. The neural network $\boldsymbol{v}_\theta$ is trained to match this velocity field by minimizing:

$$\mathcal{L}(\theta) = \mathbb{E}_{(\boldsymbol{x}_0,\boldsymbol{x}_1)\sim p_0\times p_1, t\sim\mathcal{U}(0,1)}\|\boldsymbol{v}_\theta(\boldsymbol{x}_t, t) - (\boldsymbol{x}_1 - \boldsymbol{x}_0)\|_2^2. \tag{3}$$

At inference, given $\boldsymbol{x}_0 \sim p_0$, we solve the ODE from $t = 0$ to $t = 1$ to obtain the deterministic output image $\hat{\boldsymbol{x}}_1$:

$$\hat{\boldsymbol{x}}_1 = \boldsymbol{x}_0 + \int_0^1 \boldsymbol{v}_\theta(\boldsymbol{x}_t, t)dt. \tag{4}$$

3.3 Conditional Flow Matching

Conditional flow matching (CFM) (Zhang et al., 2025b) learns a conditional vector field $\boldsymbol{v}_\theta : \mathcal{X} \times [0, 1] \times \mathcal{C} \to \mathcal{X}$ defining the flow from p_0 to a condition-dependent target distribution $p_1(\cdot|c)$ for $c \in \mathcal{C}$ via the ODE:

$$\frac{d\boldsymbol{x}_t}{dt} = \boldsymbol{v}_\theta(\boldsymbol{x}_t, t, c), \quad \boldsymbol{x}_0 \sim p_0, \boldsymbol{x}_1 \sim p_1(\cdot|c). \tag{5}$$

We apply classifier-free guidance (Ho & Salimans, 2022; Zheng et al., 2023) to enhance conditional generation. During training, we randomly mask condition c with probability p_c, replacing it with a null condition $c_\emptyset$. At inference, conditional and unconditional predictions combine to form the guided vector field:

$$\tilde{\boldsymbol{v}}_\theta(\boldsymbol{x}_t, t, c) = \alpha\boldsymbol{v}_\theta(\boldsymbol{x}_t, t, c) + (1 - \alpha)\boldsymbol{v}_\theta(\boldsymbol{x}_t, t, c_\emptyset), \tag{6}$$

where $\alpha \geq 1$ controls the guidance strength.

Henceforth, we use the conditional notation $p(\boldsymbol{x}_1|\boldsymbol{x}_0, c)$ and $p_1(\cdot|c)$ as a unified representation for both conditional and unconditional cases. The unconditional case can be recovered via $c = c_\emptyset$.

4 Methodology

We first deploy **Stochastic Flow Matching (SFM)**, which stochastically maps inputs $\boldsymbol{x}_0$ to samples $\boldsymbol{x}_1$ conditioned on c while preserving the learned transport marginals. We then formulate **Bayesian Stochastic Flow Matching (BSFM)** and propose **AVUQ** for sample-efficient nested estimation of aleatoric and epistemic uncertainty. Fig. 2 outlines our methodological components.

4.1 Stochastic Flow Matching

To quantify the aleatoric uncertainty inherent to the transformation process, a naive extension adds diffusion noise to the ODE solver of the learned velocity field:

$$d\boldsymbol{x}_t = \boldsymbol{v}_\theta(\boldsymbol{x}_t, t, c)dt + \sigma_t d\boldsymbol{W}_t, \tag{7}$$

where σ_t is a predefined noise schedule and $\boldsymbol{W}_t$ denotes standard Brownian motion. However, this directly perturbs the marginals $p_t(\boldsymbol{x}_t|c)$ at each time t, deviating from the learned transport and degrading generation quality. Following prior work (Su et al., 2025), we instead derive a marginal-preserving stochastic differential equation (SDE) via the Fokker-Planck equation. For any ODE $d\boldsymbol{x}_t = \boldsymbol{v}_\theta(\boldsymbol{x}_t, t, c)\, dt$ with marginals $\{p_t(\boldsymbol{x}_t|c)\}$, the corresponding SDE with drift correction,

$$d\boldsymbol{x}_t = \Big(\boldsymbol{v}_\theta(\boldsymbol{x}_t, t, c) - \tfrac{1}{2}\sigma_t^2 \nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t|c)\Big)dt + \sigma_t d\boldsymbol{W}_t \tag{8}$$

shares identical marginals $p_t(\boldsymbol{x}_t|c)$ for all $t \in [0, 1]$. The score correction term compensates for diffusion-induced drift, preserving the learned distribution-to-distribution mapping while introducing controlled stochasticity.

Training and Inference. We parameterize the score function as $s_\phi(\boldsymbol{x}_t, t, c) \approx \nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t|c)$ and train it jointly with the velocity field. We perturb the interpolation path following the standard score matching construction (Albergo et al., 2023):

$$\boldsymbol{x}_t = (1-t)\boldsymbol{x}_0 + t\boldsymbol{x}_1 + \gamma_t \boldsymbol{z}_t, \quad \boldsymbol{z}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \tag{9}$$

where γ_t is a smooth noise schedule with $\gamma_0 = \gamma_1 = 0$. For $\gamma_t > 0$, the score admits the analytic form $\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t|c) = -\gamma_t^{-1}\boldsymbol{z}_t$. We set $\gamma_t = a \sin^2(\pi t)$ because it vanishes at the endpoints, preserving the source and target samples, while smoothly placing the largest perturbation near the middle of the interpolation path. We minimize the combined objective $\mathcal{L}(\theta, \phi) = \mathcal{L}_{\boldsymbol{v}}(\theta) + \lambda \mathcal{L}_s(\phi)$,

$$\mathcal{L}_{\boldsymbol{v}}(\theta) = \mathbb{E}\|\boldsymbol{v}_\theta(\boldsymbol{x}_t, t, c) - (\boldsymbol{x}_1 - \boldsymbol{x}_0)\|_2^2, \quad \mathcal{L}_s(\phi) = \mathbb{E}\|s_\phi(\boldsymbol{x}_t, t, c) + \gamma_t^{-1}\boldsymbol{z}_t\|_2^2, \tag{10}$$

where expectations are taken over $(\boldsymbol{x}_0, \boldsymbol{x}_1) \sim p_0 \times p_1(\cdot|c)$ and $t \sim \mathcal{U}(0, 1)$. At inference, we solve the SDE using the learned velocity and score networks to generate samples, with $\boldsymbol{v}_\theta$ denoting the conditional or guided velocity field from the preliminaries:

$$d\boldsymbol{x}_t = \Big(\boldsymbol{v}_\theta(\boldsymbol{x}_t, t, c) - \tfrac{1}{2}\sigma_t^2 s_\phi(\boldsymbol{x}_t, t, c)\Big)dt + \sigma_t d\boldsymbol{W}_t. \tag{11}$$

4.2 Bayesian Stochastic Flow Matching

To quantify epistemic uncertainty, we adopt a Bayesian treatment of the velocity and score network parameters (θ, ϕ). Given training data $\mathcal{D}$, the posterior is given by

$$p(\theta, \phi|\mathcal{D}) \propto p(\mathcal{D}|\theta, \phi)p(\theta, \phi), \tag{12}$$

where $p(\mathcal{D}|\theta, \phi)$ denotes the likelihood and $p(\theta, \phi)$ the prior. The predictive distribution marginalizes over parameter uncertainty:

$$p(\boldsymbol{x}_1|\boldsymbol{x}_0, c, \mathcal{D}) = \iint p(\boldsymbol{x}_1|\boldsymbol{x}_0, c, \theta, \phi)p(\theta, \phi|\mathcal{D})d\theta d\phi, \tag{13}$$

where $p(\boldsymbol{x}_1|\boldsymbol{x}_0, c, \theta, \phi)$ is the conditional distribution induced by the SDE in equation 11.

4.3 Uncertainty Decomposition and Estimation

Total predictive uncertainty arises from two sources: aleatoric uncertainty, reflecting inherent stochasticity in the transformation captured by $p(\boldsymbol{x}_1|\boldsymbol{x}_0, c, \theta, \phi)$, and epistemic uncertainty, arising from limited training data and captured by the model posterior $p(\theta, \phi|\mathcal{D})$.

We quantify uncertainty via total predictive variance and decompose it using the law of total variance:

$$\mathrm{Var}(\boldsymbol{x}_1|\boldsymbol{x}_0, c, \mathcal{D}) = \underbrace{\mathbb{E}_{p(\theta,\phi|\mathcal{D})}[\mathrm{Var}(\boldsymbol{x}_1|\boldsymbol{x}_0, c, \theta, \phi)]}_{\text{aleatoric uncertainty}} + \underbrace{\mathrm{Var}_{p(\theta,\phi|\mathcal{D})}[\mathbb{E}(\boldsymbol{x}_1|\boldsymbol{x}_0, c, \theta, \phi)]}_{\text{epistemic uncertainty}}, \tag{14}$$

where the first term represents aleatoric uncertainty and the second represents epistemic uncertainty.

We employ nested sampling to estimate both uncertainty terms: for each of M posterior samples $(\theta^{(m)}, \phi^{(m)}) \sim p(\theta, \phi|\mathcal{D})$, we solve the SDE K times to generate $\boldsymbol{x}_1^{(k,m)} \sim p(\cdot|\boldsymbol{x}_0, c, \theta^{(m)}, \phi^{(m)})$ and estimate

$$\hat{\mu}^{(m)} = \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{x}_1^{(k,m)} \approx \mathbb{E}(\boldsymbol{x}_1|\boldsymbol{x}_0, c, \theta^{(m)}, \phi^{(m)}), \quad \text{and} \tag{15}$$

$$\hat{\Sigma}^{(m)} = \frac{1}{K-1}\sum_{k=1}^{K}(\boldsymbol{x}_1^{(k,m)} - \hat{\mu}^{(m)})(\boldsymbol{x}_1^{(k,m)} - \hat{\mu}^{(m)})^T \approx \mathrm{Var}(\boldsymbol{x}_1|\boldsymbol{x}_0, c, \theta^{(m)}, \phi^{(m)}). \tag{16}$$

For our estimate of aleatoric uncertainty, we compute

$$\hat{A} = \frac{1}{M}\sum_{m=1}^{M}\hat{\Sigma}^{(m)} \tag{17}$$

For epistemic uncertainty, we note that by the law of total variance (conditioning on $\boldsymbol{x}_0, c$),

$$\mathrm{Var}_{p(\theta,\phi|\mathcal{D})}[\hat{\mu}(\theta,\phi)] = \underbrace{\mathrm{Var}_{p(\theta,\phi|\mathcal{D})}\big[\mathbb{E}\big(\hat{\mu} \mid \theta,\phi\big)\big]}_{\text{epistemic uncertainty}} + \underbrace{\mathbb{E}_{p(\theta,\phi|\mathcal{D})}\big[\mathrm{Var}\big(\hat{\mu} \mid \theta,\phi\big)\big]}_{\text{residual Monte Carlo noise}}. \tag{18}$$

The first term equals the true epistemic uncertainty because $\mathbb{E}\big(\hat{\mu} \mid \theta,\phi\big) = \mathbb{E}(\boldsymbol{x}_1|\boldsymbol{x}_0, c, \theta, \phi)$. Additionally, note that $\mathrm{Var}\big(\hat{\mu} \mid \theta,\phi\big) = \frac{1}{K}\mathrm{Var}(\boldsymbol{x}_1|\boldsymbol{x}_0, c, \theta, \phi)$. That is, the residual Monte Carlo (MC) noise is aleatoric uncertainty contamination, scaled by the sample size K, that appears in the naive estimator for epistemic uncertainty:

$$\hat{E} = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{\mu}^{(m)} - \bar{\mu})(\hat{\mu}^{(m)} - \bar{\mu})^T. \tag{19}$$

Ideally, we would correct for this MC noise by subtracting $\hat{A}/K$. However, nested sampling incurs $\mathcal{O}(M \times K)$ SDE solves, which is computationally prohibitive for high-dimensional images, especially for large M and K, the regime in which our estimates are lower variance. To handle this, we introduce a sample-efficient alternative via AVUQ.

4.4 Scalable and Sample-Efficient Uncertainty Estimation

This section addresses two computational bottlenecks: sampling from an intractable neural-network posterior and estimating SDE moments with few solver calls.

AVUQ Generating nested SDE samples $\boldsymbol{x}_1^{(i,m)} \sim p(\cdot|\boldsymbol{x}_0, c, \theta^{(m)}, \phi^{(m)})$ for $i = 1, \ldots, K$ and $m = 1, \ldots, M$ is computationally expensive. To alleviate this cost, we employ antithetic sampling (Hammersley & Morton, 1956) to reduce MC variance under limited sampling budgets. Specifically, for each noise realization ϵ used to drive the SDE, we additionally simulate the trajectory induced by its negation $-\epsilon$. This construction introduces negative correlation between paired samples, reducing conditional-mean estimator variance without bias under the usual symmetry of the driving noise.

In particular, for J antithetic pairs ($K = 2J$ total samples), $\big(\boldsymbol{x}_{1,+}^{(j,m)}, \boldsymbol{x}_{1,-}^{(j,m)}\big)$, we form

$$\mathbf{y}^{(j,m)} = \frac{1}{2}\big(\boldsymbol{x}_{1,+}^{(j,m)} + \boldsymbol{x}_{1,-}^{(j,m)}\big), \quad \hat{\mu}_{anti}^{(m)} = \frac{1}{J}\sum_{j=1}^{J}\mathbf{y}^{(j,m)}. \tag{20}$$

We use $\hat{\mu}_{anti}^{(m)}$ as the conditional-mean estimate in the epistemic estimator. For aleatoric uncertainty, we compute the within-posterior-sample marginal variance from the raw antithetic trajectories $\{\boldsymbol{x}_{1,+}^{(j,m)}, \boldsymbol{x}_{1,-}^{(j,m)}\}_{j=1}^{J}$, not from the pair means. Importantly, antithetic sampling reduces the residual MC error in equation 18, enabling more reliable epistemic uncertainty estimation under tight computational budgets. See Fig. 2.

Estimating the Model Posterior $p(\theta, \phi \mid \mathcal{D})$**.** AVUQ relies on posterior samples $\theta^{(m)}, \phi^{(m)}$. Exact Bayesian inference for deep neural networks is typically intractable. We therefore adopt MC-Dropout, a scalable approach to approximate the posterior (Gal & Ghahramani, 2016). As shown in Fig. 2, we insert dropout layers in the velocity and score networks. At inference, we keep dropout active and perform multiple stochastic forward passes, yielding samples $\{(\theta^{(m)}, \phi^{(m)})\}_{m=1}^{M}$ from the approximate posterior. This provides a simple and scalable method in high-dimensional parameter spaces. We note that AVUQ works with any approximate posterior sampling method. See Sec. 5.3 for empirical comparisons. Here we focus on using AVUQ for unreliable generation detection. Assessing conditional calibration is challenging without matched source–target observations in this high-dimensional, unpaired setting; we leave this evaluation to future work.

AVUQ for unreliable generation detection. For our scalar anomaly scores, $f(\boldsymbol{x}_0, c)$, we compute $-\mathrm{tr}(\hat{E})$ and $-\mathrm{tr}(\hat{A})$ for epistemic and aleatoric uncertainty, respectively. These scores are traces, so we never form or store full pixel-level covariance matrices. Since paired trajectories are correlated, this aleatoric trace is used as a practical scalar ranking score rather than an unbiased independent-trajectory variance estimator. For MAP, we use the negative mean pixel-wise variance as the anomaly score. We then apply a binary decision rule:

$$h(\boldsymbol{x}_0, c) = \begin{cases} \text{Unreliable}, & \text{if } f(\boldsymbol{x}_0, c) > \tau \\ \text{Reliable}, & \text{if } f(\boldsymbol{x}_0, c) \leq \tau \end{cases} \tag{21}$$

where τ is a threshold parameter. The sign convention reflects the failure mode we observe in these generative models: under distribution shift, the model can collapse toward overconfident predictions, so uncertainty may decrease rather than increase on OOD inputs (Kirichenko et al., 2020; Zhang et al., 2021). In this regime, unusually low epistemic uncertainty is itself an anomaly signal, motivating $-\mathrm{tr}(\hat{E})$ rather than $\mathrm{tr}(\hat{E})$. We apply the same convention to aleatoric uncertainty because it also decreases under the shifts considered in Sec. 5.3 and is particularly sensitive to mild deviations. This direction should be revalidated in settings where OOD inputs instead induce genuine posterior disagreement rather than collapse.

5 Results

We assess SFM for generalization and BSFM with AVUQ for unreliable generation detection on cellular imaging and fMRI tasks, comparing against a range of baseline alternatives.

5.1 Datasets

We evaluate on two cell-imaging benchmarks: **BBBC021** (chemical perturbations) (Caie et al., 2010) and **JUMP** (chemical and genetic perturbations) (Chandrasekaran et al., 2023). To probe robustness and unreliable generation detection, we define four scenarios spanning mild to severe distribution shifts: **Unseen Plates** and **Unseen Cell Lines** in JUMP, synthetic **Intensity Shift** in BBBC021 to mimic imaging-condition variation across laboratory settings, and a combined **Unseen Pert.** setting with unseen perturbations and intensity shifts. We further evaluate **Intensity Shift** scenarios in JUMP for generation quality and unreliable generation detection, spanning a full range of distribution shifts from mild to severe.

To further assess generalization performance, we additionally consider the **fMRI** Theory-of-Mind (ToM) dataset (Richardson et al., 2018), where the objective is to transform resting-state scans to task-activated states. We train on high-performing ToM subjects and evaluate generalization to low-performing (LToM) subjects. Table 3 summarizes datasets and scenarios. More details are provided in Appendix B.

5.2 Experimental Setup

Baselines. For generalization, we compare **SFM** against: **CellFlux** (Zhang et al., 2025b), a state-of-the-art deterministic flow matching model for distribution-to-distribution cell image generation; **BBDM**, **GOUB**, and **UniDB**, three diffusion-bridge baselines (Li et al., 2023a; Yue et al., 2023; Zhu et al., 2025); **UNSB** (Kim et al., 2023), which employs a multi-step GAN to learn a Schrödinger bridge between source and target distributions; and **SDEdit** (Meng et al., 2021), which performs partial noising of the source sample followed by denoising to generate the target sample. For unreliable generation detection, we establish a benchmark by comparing different UQ approaches within the distribution-to-distribution flow matching framework. We compare **AVUQ** to i.i.d. nested sampling of BSFM with **MC-Dropout (MCD)** (Gal & Ghahramani, 2016), **Laplace Approximation (LA)** (Kou et al., 2023), and **Stochastic Weight Averaging-Gaussian (SWAG)** (Maddox et al., 2019). We also include two baselines that do not require nested sampling to provide lower-cost points in the cost-performance tradeoff: a **MAP** baseline that approximates aleatoric uncertainty using only the MAP estimate (see Appendix A) and requires only $\mathcal{O}(K)$ SDE solves, and a **MCD-DFM** baseline that applies MC-Dropout to a Bayesian deterministic flow matching model without the SFM stochastic extension and requires only $\mathcal{O}(M)$ posterior samples to approximate epistemic uncertainty.

Table 1: Generalization performance. We report FID and KID scores across six unseen scenarios on BBBC021, JUMP, and fMRI datasets. Best results are in **bold**, second best are underlined.

Method	BBBC021				JUMP						fMRI	
	Unseen Pert.		Intensity Shift		Unseen Cell Lines		Unseen Plates		Intensity Shift		Low ToM	
	FID↓	KID↓	FID↓	KID↓	FID↓	KID↓	FID↓	KID↓	FID↓	KID↓	FID↓	KID↓
BBDM	107.69	11.88	86.34	9.25	126.41	13.25	107.24	10.68	133.81	16.07	45.31	3.55
GOUB	51.89	4.71	36.94	3.46	62.47	6.62	57.66	6.14	126.51	14.49	33.49	2.63
UniDB	50.16	4.33	36.78	3.06	65.1	6.9	60.33	6.34	121.33	13.91	30.23	1.85
UNSB	88.31	6.27	56.95	4.7	57.24	4.19	45.5	4.04	115.42	12.74	75.54	8.89
SDEdit	37.18	3.48	29.57	2.72	33.96	3.11	**13.21**	**0.9**	122.23	13.32	56.87	6.15
CellFlux	103.73	12.76	62.01	5.84	39.06	3.4	17.81	1.04	209.66	25.06	34.86	2.63
SFM	**33.29**	**2.02**	**28.14**	**1.87**	**25.1**	**1.75**	18.02	1.23	**107.60**	**10.53**	**25.55**	**1.47**

Table 2: Unreliable generation detection performance. We report AUROC and AUPR scores as mean ± standard error over 3 random seeds across 4 unreliable-generation scenarios on BBBC021 and JUMP datasets. We group methods into non-nested baselines, nested i.i.d. baselines, and nested-antithetic AVUQ.

Method	BBBC021				JUMP			
	Unseen Pert.		Intensity Shift		Unseen Cell Lines		Intensity Shift	
	AUROC↑	AUPR↑	AUROC↑	AUPR↑	AUROC↑	AUPR↑	AUROC↑	AUPR↑
non-nested								
MAP Aleatoric	0.691±0.033	0.482±0.063	0.711±0.018	0.637±0.023	0.799±0.021	0.746±0.024	0.276±0.07	0.325±0.017
MCD-DFM Epistemic	0.621±0.056	0.332±0.047	0.654±0.051	0.532±0.05	0.645±0.03	0.545±0.02	0.425±0.047	0.368±0.017
nested-i.i.d.								
SWAG Aleatoric	0.662±0.044	0.472±0.076	0.627±0.059	0.610±0.052	0.855±0.005	0.816±0.004	0.357±0.080	0.35±0.024
SWAG Epistemic	0.676±0.063	0.499±0.099	0.645±0.064	0.639±0.06	0.832±0.001	0.800±0.005	0.573±0.071	0.444±0.038
LA Aleatoric	0.666±0.033	0.474±0.051	0.636±0.048	0.623±0.041	0.865±0.003	0.832±0.001	0.324±0.056	0.338±0.016
LA Epistemic	0.644±0.023	0.422±0.043	0.657±0.058	0.648±0.04	0.829±0.004	0.787±0.002	0.503±0.043	0.411±0.023
MCD Aleatoric	0.741±0.004	0.501±0.011	0.72±0.009	0.645±0.015	0.856±0.01	0.822±0.012	0.438±0.005	0.376±0.002
MCD Epistemic	0.731±0.012	0.486±0.014	0.729±0.01	0.642±0.012	0.833±0.022	0.796±0.029	0.418±0.009	0.368±0.004
nested-antithetic								
AVUQ Aleatoric	0.742±0.018	0.523±0.06	0.729±0.03	0.663±0.04	**0.869±0.007**	**0.835±0.006**	0.309±0.072	0.335±0.02
AVUQ Epistemic	**0.809±0.019**	**0.698±0.059**	**0.785±0.026**	**0.768±0.033**	0.794±0.006	0.726±0.01	**0.89±0.005**	**0.799±0.014**

Training and evaluation details. We evaluate generation quality using Fréchet Inception Distance (FID) and Kernel Inception Distance (KID) to measure distribution similarity. For unreliable generation detection, we first take all scenarios unseen during training and filter them to account for the fact that SFM can generalize well to some unseen scenarios, as demonstrated in Table 1. We filter samples using prediction error measured in the feature space of a mode-of-action pretrained classifier (BBBC021) (Zhang et al., 2025b) or in Structural Similarity Index Measure (SSIM) space (JUMP), and flag only high-error cases as unreliable generations. Further details are in Appendix C. Appendix D reports sensitivity analyses showing that the main unreliable generation detection trends are robust to smaller sampling budgets and alternative scenario-filtering threshold choices. We also include OOD provenance detection, where we do not filter unseen scenarios at all.

Following CellFlux (Zhang et al., 2025b), we adopt a U-Net-based architecture to parameterize the velocity field. Perturbation conditions are encoded using IMPA (Palma et al., 2025). We train models for 100 epochs on BBBC021, and 200 epochs on JUMP and fMRI using two NVIDIA H100 GPUs. For generation quality, we evaluate using 5000 generated images per unseen scenario on BBBC021 and JUMP, and 1024 images on fMRI. For unreliable generation detection, we evaluate on randomly selected 500 ID and 500 OOD images per scenario (except 250 OOD images for Unseen Pert due to data availability). For MAP and MCD-DFM, we generate 4 samples per image. For AVUQ , MCD, LA, and SWAG, we generate 4×4 samples per image via nested sampling, where the 4 SDE samples for AVUQ come from 2 antithetic pairs.

5.3 Main Results

Stochastic Flow Matching Improves Generalization. As shown in Table 1, SFM demonstrates significant and consistent improvements in generalization across all scenarios except Unseen Plates, where the distribution shift is the smallest and all baselines perform well. On JUMP Intensity Shift, SFM achieves FID/KID of 107.60/10.53, compared with 115.42/12.74 for

the next-best UNSB. This overall improved robustness stems from explicitly modeling aleatoric uncertainty through stochastic perturbations during training and SDE-based sampling at inference. By capturing the inherent variability in the conditional distribution $p(\boldsymbol{x}_1 \mid \boldsymbol{x}_0, c, \theta, \phi)$[1] rather than learning a deterministic mapping, the model avoids overfitting to spurious correlations specific to the training distribution. The resulting uncertainty-aware generation process produces diverse, plausible outputs that better generalize to unseen distribution shifts. Controlled training and inference ablations appear in Appendix F.1. Examples of the generated images together with the corresponding source and ground truth target images are shown in Fig. 3.

Unreliable Generation Detection Performance Analysis. As shown in Table 2, uncertainty-based scores provide informative signals for unreliable generation detection across cellular imaging shifts. Among non-nested baselines, MAP provides a strong lower-cost aleatoric baseline in several cases; however, it significantly underperforms AVUQ in all scenarios. MCD-DFM is generally weaker than nested-i.i.d. BSFM baselines, indicating that the SFM stochastic backbone is important for UQ. AVUQ provides the most consistent uncertainty signals for unreliable generation detection, demonstrating the importance of combining SFM with antithetic sampling. In some scenarios, epistemic uncertainty provides the stronger signal, whereas in other cases aleatoric uncertainty is critical, indicating that both signals are needed for robust unreliable generation detection. We omit JUMP Unseen Plates because, as shown in Table 1, SFM produces reliable generations for nearly all samples under this mild distribution shift.

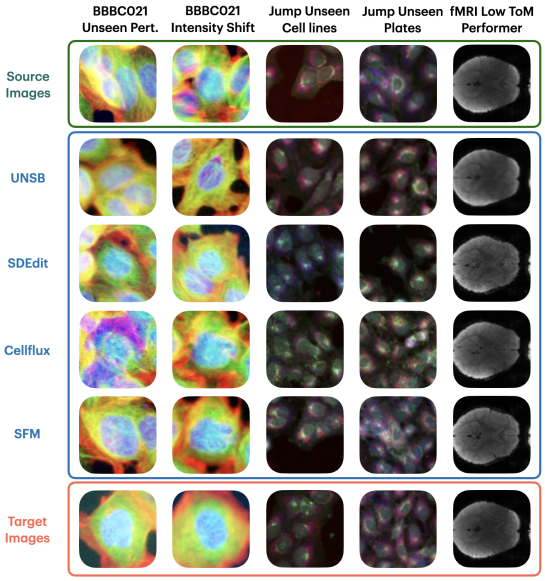


Figure 3: Example generated images from different methods under various unseen scenarios, compared with source and ground-truth target images.

Ablation Study. We further ablate two practical factors in unreliable generation detection: sampling budget and the construction of unreliable-generation labels. A key challenge for nested UQ is that finite-sample aleatoric variability can contaminate epistemic estimates, making the two signals difficult to use separately. Antithetic sampling reduces this residual Monte Carlo error in AVUQ, which helps separate the two uncertainty components under small sampling budgets. Table 4 shows that the main trends are already preserved with small budgets such as 2×4 or 4×2, supporting the sample-efficiency of the estimator. We also test whether the results depend on the scenario-filtering threshold used to define unreliable generations. Tables 5 and 6 show that the relative trends are broadly stable when using no thresholding or a mean threshold, respectively, indicating that the conclusions are not driven by a single label-construction choice. Direct equal-budget variance results appear in Appendix F.2.

6 Conclusion and Limitations

We showed how SFM improves generalization in distribution-to-distribution scientific imaging and introduced AVUQ for efficient estimation of aleatoric and epistemic uncertainty in BSFM using antithetic sampling. Scalar scores derived from these estimates support unreliable generation detection. Limitations include approximate posterior inference, numerical SDE integration, and trace-based covariance scores. Future work will evaluate these components and validate the anomaly-score direction across broader shifts.

[1] In principle, integrating over posterior uncertainty in (θ, ϕ) could further improve generalization; however, averaging generated images $\boldsymbol{x}_1$ pixel-wise reduces image quality.

AI USE STATEMENT

In preparing this paper, we used generative AI tools to assist with LaTeX template formatting, manuscript editing, and the organization and presentation of experimental results in the main text and appendix. The authors reviewed the AI-assisted revisions and take responsibility for the final manuscript and accompanying artifacts.

REFERENCES

Lemar Abdi, Amaan Valiuddin, Francisco Caetano, Christiaan Viviers, and Fons van der Sommen. Zero-shot image anomaly detection using generative foundation models. *arXiv preprint arXiv:2507.22692*, 2025.

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.

Fuat Arslan, Bilal Kabas, Onat Dalmaz, Muzaffer Ozbey, and Tolga Çukur. Self-consistent recursive diffusion bridge for medical image translation. *Medical Image Analysis*, 106:103747, 2025.

Cosmin I Bercea, Benedikt Wiestler, Daniel Rueckert, and Julia A Schnabel. Diffusion models with implicit guidance for medical anomaly detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 211–220. Springer, 2024.

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.

Wolfgang Blau, Vinton G Cerf, Juan Enriquez, Joseph S Francisco, Urs Gasser, Mary L Gray, Mark Greaves, Barbara J Grosz, Kathleen Hall Jamieson, Gerald H Haug, et al. Protecting scientific integrity in an age of generative ai, 2024.

Peter D Caie, Rebecca E Walls, Alexandra Ingleston-Orme, Sandeep Daya, Tom Houslay, Rob Eagle, Mark E Roberts, and Neil O Carragher. High-content phenotypic profiling of drug response signatures across distinct cancer cells. *Molecular cancer therapeutics*, 9(6):1913–1926, 2010.

Matthew Chan, Maria Molina, and Chris Metzler. Estimating epistemic and aleatoric uncertainty with a single model. *Advances in Neural Information Processing Systems*, 37:109845–109870, 2024.

Srinivas Niranj Chandrasekaran, Jeanelle Ackerman, Eric Alix, D Michael Ando, John Arevalo, Melissa Bennion, Nicolas Boisseau, Adriana Borowa, Justin D Boyd, Laurent Brino, et al. Jump cell painting dataset: morphological impact of 136,000 chemical and genetic perturbations. *BioRxiv*, pp. 2023–03, 2023.

Sungik Choi, Hankook Lee, Honglak Lee, and Moontae Lee. Projection regret: Reducing background bias for novelty detection via diffusion models. *Advances in Neural Information Processing Systems*, 36:19230–19245, 2023.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in neural information processing systems*, 34:17695–17709, 2021.

Michele De Vita and Vasileios Belagiannis. Diffusion model guided sampling with pixel-wise aleatoric uncertainty estimation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3844–3854. IEEE, 2025.

Canberk Ekmekci and Mujdat Cetin. Conformalized generative bayesian imaging: An uncertainty quantification framework for computational imaging. *arXiv preprint arXiv:2504.07696*, 2025.

Berthy T Feng, Jamie Smith, Michael Rubinstein, Huiwen Chang, Katherine L Bouman, and William T Freeman. Score-based diffusion models as principled priors for inverse imaging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10520–10531, 2023.

Gianni Franchi, Nacim Belkhir, Dat Nguyen Trong, Guoxuan Xia, and Andrea Pilzer. Towards understanding and quantifying uncertainty for text-to-image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8062–8072, 2025.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Silvio Galesso, Philipp Schröppel, Hssan Driss, and Thomas Brox. Diffusion for out-of-distribution detection on road scenes and beyond. In *European Conference on Computer Vision*, pp. 110–126. Springer, 2024.

Ruiyuan Gao, Chenchen Zhao, Lanqing Hong, and Qiang Xu. Diffguard: Semantic mismatch-guided out-of-distribution detection using pre-trained diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1579–1589, 2023.

Mark S Graham, Walter HL Pinaya, Petru-Daniel Tudosiu, Parashkev Nachev, Sebastien Ourselin, and Jorge Cardoso. Denoising diffusion models for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2948–2957, 2023a.

Mark S Graham, Walter Hugo Lopez Pinaya, Paul Wright, Petru-Daniel Tudosiu, Yee H Mah, James T Teo, H Rolf Jäger, David Werring, Parashkev Nachev, Sebastien Ourselin, et al. Unsupervised 3d out-of-distribution detection with latent diffusion models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 446–456. Springer, 2023b.

John Michael Hammersley and Keith William Morton. A new monte carlo technique: antithetic variates. In *Mathematical proceedings of the Cambridge philosophical society*, volume 52, pp. 449–475. Cambridge University Press, 1956.

Alvin Heng, Harold Soh, et al. Out-of-distribution detection with a single unconditional diffusion model. *Advances in Neural Information Processing Systems*, 37:43952–43974, 2024.

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

Yue Huang, Chujie Gao, Siyuan Wu, Haoran Wang, Xiangqi Wang, Yujun Zhou, Yanbo Wang, Jiayi Ye, Jiawen Shi, Qihui Zhang, et al. On the trustworthiness of generative foundation models: Guideline, assessment, and perspective. *arXiv preprint arXiv:2502.14296*, 2025.

Metod Jazbec, Eliot Wong-Toi, Guoxuan Xia, Dan Zhang, Eric Nalisnick, and Stephan Mandt. Generative uncertainty in diffusion models. *arXiv preprint arXiv:2502.20946*, 2025.

Xuan Kan, Hejie Cui, Joshua Lukemire, Ying Guo, and Carl Yang. Fbnetgen: Task-aware gnn-based fmri analysis via functional brain network generation. In *International conference on medical imaging with deep learning*, pp. 618–637. PMLR, 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.

Beomsu Kim, Gihyun Kwon, Kwanyoung Kim, and Jong Chul Ye. Unpaired image-to-image translation via neural schr\" odinger bridge. *arXiv preprint arXiv:2305.15086*, 2023.

Eunwoo Kim, Un Yang, Cheol Lae Roh, and Stefano Ermon. Unsupervised anomaly detection using diffusion trend analysis for display inspection. *arXiv preprint arXiv:2407.09578*, 2024.

Polina Kirichenko, Pavel Izmailov, and Andrew G Wilson. Why normalizing flows fail to detect out-of-distribution data. *Advances in neural information processing systems*, 33:20578–20589, 2020.

Siqi Kou, Lei Gan, Dequan Wang, Chongxuan Li, and Zhijie Deng. Bayesdiff: Estimating pixel-wise uncertainty in diffusion via bayesian inference. *arXiv preprint arXiv:2310.11142*, 2023.

Junbeom Kwon, Jungwoo Seo, Heehwan Wang, Taesup Moon, Shinjae Yoo, and Jiook Cha. Predicting task-related brain activity from resting-state brain dynamics with fmri transformer. *Imaging Neuroscience*, 3:imag_a_00440, 2025.

Georges Le Bellier and Nicolas Audebert. Detecting out-of-distribution earth observation images with diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 481–491. IEEE, 2024.

Bo Li, Kaitao Xue, Bin Liu, and Yu-Kun Lai. Bbdm: Image-to-image translation with brownian bridge diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1952–1961. IEEE, 2023a.

Yunxiang Li, Hua-Chieh Shao, Xiao Liang, Liyuan Chen, Ruiqi Li, Steve Jiang, Jing Wang, and You Zhang. Zero-shot medical image translation via frequency-guided diffusion models. *IEEE transactions on medical imaging*, 43(3):980–993, 2023b.

Jasper Linmans, Gabriel Raya, Jeroen van der Laak, and Geert Litjens. Diffusion models for out-of-distribution detection in digital pathology. *Medical Image Analysis*, 93:103088, 2024.

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

Yaron Lipman, Marton Havasi, Peter Holderrieth, Neta Shaul, Matt Le, Brian Karrer, Ricky TQ Chen, David Lopez-Paz, Heli Ben-Hamu, and Itai Gat. Flow matching guide and code. *arXiv preprint arXiv:2412.06264*, 2024.

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.

Zhenzhen Liu, Jin Peng Zhou, Yufan Wang, and Kilian Q Weinberger. Unsupervised out-of-distribution detection with diffusion inpainting. In *International Conference on Machine Learning*, pp. 22528–22538. PMLR, 2023.

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.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

Alessandro Palma, Fabian J Theis, and Mohammad Lotfollahi. Predicting cell morphological responses to perturbations using generative modeling. *Nature Communications*, 16(1):505, 2025.

Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European conference on computer vision*, pp. 319–345. Springer, 2020.

Hilary Richardson, Grace Lisandrelli, Alexa Riobueno-Naylor, and Rebecca Saxe. Development of the social brain from age three to twelve years. *Nature communications*, 9(1):1027, 2018.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14318–14328. IEEE, 2022.

Yuyang Shi, Valentin De Bortoli, Andrew Campbell, and Arnaud Doucet. Diffusion schrödinger bridge matching. *Advances in Neural Information Processing Systems*, 36:62183–62223, 2023.

Shirin Shoushtari, Yi Wang, Xiao Shi, M Salman Asif, and Ulugbek S Kamilov. Eigenscore: Ood detection using covariance in diffusion models. *arXiv preprint arXiv:2510.07206*, 2025.

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*, 2020.

Shiye Su, Yuhui Zhang, Linqi Zhou, Rajesh Ranganath, and Serena Yeung-Levy. Three forms of stochastic injection for improved distribution-to-distribution generative modeling. *arXiv preprint arXiv:2510.06634*, 2025.

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.

Zihui Wu, Yu Sun, Yifan Chen, Bingliang Zhang, Yisong Yue, and Katherine Bouman. Principled probabilistic imaging using diffusion models as plug-and-play priors. *Advances in Neural Information Processing Systems*, 37:118389–118427, 2024.

Yutong Xie and Quanzheng Li. Measurement-conditioned denoising diffusion probabilistic model for under-sampled medical image reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 655–664. Springer, 2022.

Conghan Yue, Zhengwei Peng, Junlong Ma, Shiyan Du, Pengxu Wei, and Dongyu Zhang. Image restoration through generalized ornstein-uhlenbeck bridge. *arXiv preprint arXiv:2312.10299*, 2023.

Hui Zhang, Zheng Wang, Dan Zeng, Zuxuan Wu, and Yu-Gang Jiang. Diffusionad: Norm-guided one-step denoising diffusion for anomaly detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025a.

Lily Zhang, Mark Goldstein, and Rajesh Ranganath. Understanding failures in out-of-distribution detection with deep generative models. In *International Conference on Machine Learning*, pp. 12427–12436. PMLR, 2021.

Yuhui Zhang, Yuchang Su, Chenyu Wang, Tianhong Li, Zoe Wefers, Jeffrey Nirschl, James Burgess, Daisy Ding, Alejandro Lozano, Emma Lundberg, et al. Cellflux: Simulating cellular morphology changes via flow matching. *arXiv preprint arXiv:2502.09775*, 2025b.

Qinqing Zheng, Matt Le, Neta Shaul, Yaron Lipman, Aditya Grover, and Ricky TQ Chen. Guided flows for generative modeling and decision making. *arXiv preprint arXiv:2311.13443*, 2023.

Linqi Zhou, Aaron Lou, Samar Khanna, and Stefano Ermon. Denoising diffusion bridge models. *arXiv preprint arXiv:2309.16948*, 2023.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

Kaizhen Zhu, Mokai Pan, Yuexin Ma, Yanwei Fu, Jingyi Yu, Jingya Wang, and Ye Shi. Unidb: A unified diffusion bridge framework via stochastic optimal control. *arXiv preprint arXiv:2502.05749*, 2025.

A MAP APPROXIMATION FOR ALEATORIC UNCERTAINTY

Proposition A.1 (MAP Approximation for Aleatoric Uncertainty)**.** *Let* $V(\theta, \phi) := \mathrm{tr}(\mathrm{Var}(\boldsymbol{x}_1|\boldsymbol{x}_0, c, \theta, \phi))$ *be the scalar aleatoric uncertainty score used downstream. Let* $q(\theta, \phi) = \mathcal{N}((\hat{\theta}, \hat{\phi}), \boldsymbol{H}^{-1})$ *be a Laplace surrogate centered at the MAP estimate. If* V *is* L*-Lipschitz continuous in* (θ, ϕ) *and* $\|\boldsymbol{H}^{-1}\|_{\mathrm{op}} \leq \epsilon^2$*, then*

$$\left|\mathbb{E}_{q(\theta,\phi)}[V(\theta, \phi)] - V(\hat{\theta}, \hat{\phi})\right| \leq L\epsilon\sqrt{d}, \tag{22}$$

where $d = \dim(\theta, \phi)$.

Remark A.2. The bound is a surrogate-local statement and can be loose in high dimensions. It justifies the MAP plug-in as a computationally cheap approximation under posterior concentration, but it does not imply that MAP is empirically optimal; in our experiments, nested sampling can provide stronger unreliable generation detection when the additional SDE solves are affordable.

Proof. We prove that the MAP estimate provides a good approximation to the expected scalar aleatoric uncertainty score under the Laplace surrogate stated in Proposition A.1.

Setup. Let $V(\theta, \phi) := \mathrm{tr}(\mathrm{Var}(\boldsymbol{x}_1|\boldsymbol{x}_0, c, \theta, \phi))$ denote the scalar conditional uncertainty score as a function of parameters. Let $q(\theta, \phi) = \mathcal{N}((\hat{\theta}, \hat{\phi}), \boldsymbol{H}^{-1})$, where $(\hat{\theta}, \hat{\phi})$ is the MAP estimate and $\boldsymbol{H}$ is the Hessian of the negative log-posterior at the MAP.

Step 1: Apply Lipschitz continuity. By the triangle inequality and Lipschitz continuity of V:

$$\begin{aligned}
&\left|\mathbb{E}_{q(\theta,\phi)}[V(\theta, \phi)] - V(\hat{\theta}, \hat{\phi})\right| \\
&= \left|\mathbb{E}_{q(\theta,\phi)}[V(\theta, \phi) - V(\hat{\theta}, \hat{\phi})]\right| \\
&\leq \mathbb{E}_{q(\theta,\phi)}[|V(\theta, \phi) - V(\hat{\theta}, \hat{\phi})|] \\
&\leq L \cdot \mathbb{E}_{q(\theta,\phi)}[\|(\theta, \phi) - (\hat{\theta}, \hat{\phi})\|_2],
\end{aligned} \tag{23}$$

where the last inequality uses the L-Lipschitz property: $|V(\theta, \phi) - V(\theta', \phi')| \leq L\|(\theta, \phi) - (\theta', \phi')\|_2$.

Step 2: Bound the expected deviation. Let $\delta := (\theta, \phi) - (\hat{\theta}, \hat{\phi}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{H}^{-1})$ denote the parameter deviation from the MAP. By Jensen's inequality (since $\|\cdot\|_2$ is convex):

$$\mathbb{E}[\|\delta\|_2] = \mathbb{E}\left[\sqrt{\sum_{i=1}^{d} \delta_i^2}\right] \leq \sqrt{\mathbb{E}\left[\sum_{i=1}^{d} \delta_i^2\right]} = \sqrt{\mathrm{tr}(\boldsymbol{H}^{-1})}, \tag{24}$$

where $d = \dim(\theta, \phi)$ is the total parameter dimension.

Step 3: Apply the operator norm bound. Using the relationship between trace and operator norm:

$$\mathrm{tr}(\boldsymbol{H}^{-1}) = \sum_{i=1}^{d} \lambda_i(\boldsymbol{H}^{-1}) \leq d \cdot \max_i \lambda_i(\boldsymbol{H}^{-1}) = d \cdot \|\boldsymbol{H}^{-1}\|_{\mathrm{op}}, \tag{25}$$

where $\lambda_i(\boldsymbol{H}^{-1})$ are the eigenvalues of $\boldsymbol{H}^{-1}$ and $\|\boldsymbol{H}^{-1}\|_{\mathrm{op}}$ is the operator norm.

Step 4: Combine the bounds. By assumption, $\|\boldsymbol{H}^{-1}\|_{\mathrm{op}} \leq \epsilon^2$. Therefore:

$$\mathbb{E}[\|\delta\|_2] \leq \sqrt{\mathrm{tr}(\boldsymbol{H}^{-1})} \leq \sqrt{d \cdot \|\boldsymbol{H}^{-1}\|_{\mathrm{op}}} \leq \sqrt{d \cdot \epsilon^2} = \epsilon\sqrt{d}. \tag{26}$$

Step 5: Final bound. Combining Steps 1 and 4:

$$\left|\mathbb{E}_{q(\theta,\phi)}[V(\theta, \phi)] - V(\hat{\theta}, \hat{\phi})\right| \leq L \cdot \mathbb{E}[\|\delta\|_2] \leq L\epsilon\sqrt{d}. \tag{27}$$

□

Table 3: Summary of datasets and OOD scenarios.

Dataset	BBBC021	BBBC021	JUMP	JUMP	JUMP	fMRI
OOD Case	Unseen Pert.	Intensity Shift	Unseen Cell Lines	Unseen Plates	Intensity Shift	Low ToM
# Channels	3	3	5	5	5	1
# Images	98K	98K	72K	72K	72K	12K
Dist Shift	Extreme	High	Medium	Low	High	-

B Dataset and OOD Scenario Details

Cell Imaging Datasets. BBBC021. The BBBC021v1 dataset (Caie et al., 2010) from the Broad Bioimage Benchmark Collection is a benchmark for image-based phenotypic profiling of chemical perturbations in MCF-7 breast cancer cells. It contains 97,504 fluorescent microscopy images captured from cells treated with 113 small molecules at eight concentrations, targeting diverse cellular mechanisms including actin disruption, Aurora kinase inhibition, and microtubule stabilization. Each image provides multi-channel fluorescence for DNA, F-actin, and beta-tubulin, enabling detailed morphological analysis.

JUMP. The JUMP dataset (Chandrasekaran et al., 2023) is the most comprehensive image-based profiling resource to date, integrating both genetic and chemical perturbations. It comprises approximately 3 million images capturing phenotypic responses of 75 million single cells to genetic knockouts (CRISPR/ORF) and chemical treatments. The dataset includes two cell types: U2OS and A549. We use A549 cells for model training and U2OS cells for OOD evaluation. For our experiments, we use a subset of 72,000 images and focus exclusively on chemical perturbations.

fMRI Dataset. We use the ds000228-1.1.1 dataset (Richardson et al., 2018) containing MRI data from 3–12-year-old children viewing a Pixar animated film. The task is to transform brain states from resting baseline (TRs 1–10, 20 seconds) to Theory of Mind (ToM) activation (7 events totaling 25 TRs, 50 seconds). We focus on slices 13–16 (4 middle brain slices capturing ToM-responsive regions: temporoparietal junction and medial prefrontal cortex), with each slice as a 64×64 grayscale image. The dataset includes 122 children categorized by False Belief performance: PASS group (84 subjects, mean ToM score: 0.883 ± 0.102, age: 7.6 ± 2.2 years) as in-domain, and INC+FAIL groups (38 subjects, mean ToM score: 0.537 ± 0.148, age: 4.7 ± 1.0 years) as out-of-distribution. Each subject contributes 40 rest images and 100 ToM images. The training set contains 9,712 images from PASS subjects; the test ID set contains 2,048 randomly selected PASS images; and the test OOD set contains 2,048 randomly selected INC+FAIL images, all normalized per slice.

C Implementation Details

Filtering Strategy for Unreliable Generation Detection. To ensure the reliability of our unreliable generation detection benchmarks, we filter out ambiguous samples that may not represent true in-distribution (ID) or out-of-distribution (OOD) behavior. For BBBC021, we compute the feature-space distance between model predictions and randomly paired ground-truth treatment images using a pretrained mode-of-action classifier. For JUMP, we use the Structural Similarity Index Measure (SSIM) as the distance metric. We apply thresholds based on the distances calculated across the dataset. Specifically, we remove ID samples with distances greater than $\mu - 0.5\sigma$ (potentially poor generations) and remove OOD samples with distances smaller than $\mu + 0.5\sigma$ (samples that appear misleadingly in-distribution), where μ and σ denote the mean and standard deviation of the distance metric, respectively.

D Unreliable Generation Detection Sensitivity Analyses

Table 4 reports sensitivity to the number of posterior samples and SDE samples used by AVUQ. The results show that even small sampling budgets, such as 2×4 or 4×2, already preserve the main unreliable generation detection trends, supporting the sample-efficiency role of antithetic sampling.

Method	BBBC021				JUMP					
	Unseen Pert.		Intensity Shift		Unseen Cell Lines		Unseen Plates		Intensity Shift	
	AUROC↑	AUPR↑	AUROC↑	AUPR↑	AUROC↑	AUPR↑	AUROC↑	AUPR↑	AUROC↑	AUPR↑
2x2 Aleatoric	0.673	0.404	0.679	0.572	**0.838**	**0.792**	**0.735**	**0.662**	0.303	0.331
2x2 Epistemic	**0.752**	**0.551**	**0.785**	**0.715**	0.743	0.678	0.628	0.595	**0.826**	**0.710**
2x4 Aleatoric	0.718	0.469	0.715	0.647	**0.852**	**0.818**	**0.747**	**0.673**	0.363	0.349
2x4 Epistemic	**0.725**	**0.597**	**0.784**	**0.763**	0.759	0.709	0.614	0.542	**0.846**	**0.737**
4x2 Aleatoric	0.694	0.426	0.688	0.598	**0.856**	**0.813**	**0.743**	**0.669**	0.278	0.323
4x2 Epistemic	**0.773**	**0.621**	**0.784**	**0.736**	0.779	0.709	0.671	0.594	**0.891**	**0.801**
4x4 Aleatoric	0.730	0.486	0.720	0.649	**0.860**	**0.829**	**0.739**	**0.666**	0.364	0.350
4x4 Epistemic	**0.788**	**0.648**	**0.795**	**0.766**	0.800	0.738	0.667	0.578	**0.897**	**0.816**
8x4 Aleatoric	0.730	0.500	0.709	0.643	**0.870**	**0.843**	**0.744**	**0.676**	0.357	0.347
8x4 Epistemic	**0.807**	**0.672**	**0.770**	**0.750**	0.803	0.745	0.679	0.599	**0.906**	**0.831**

Table 4: Sample-size sensitivity for AVUQ. Each row reports AUROC/AUPR for a posterior-sample $\times$ SDE-sample budget.

Method	BBBC021				JUMP					
	Unseen Pert.		Intensity Shift		Unseen Cell Lines		Unseen Plates		Intensity Shift	
	AUROC↑	AUPR↑	AUROC↑	AUPR↑	AUROC↑	AUPR↑	AUROC↑	AUPR↑	AUROC↑	AUPR↑
MAP Aleatoric	0.616	0.443	0.636	0.626	0.670	0.627	0.517	0.519	0.227	0.351
SWAG Aleatoric	0.684	0.589	0.680	0.729	0.699	0.647	0.522	0.520	0.813	0.712
SWAG Epistemic	0.661	0.572	0.680	0.730	0.689	0.638	0.516	0.518	0.806	0.702
LA Aleatoric	0.662	0.574	0.671	0.727	0.684	0.646	0.527	0.528	0.804	0.707
LA Epistemic	0.635	0.495	0.669	0.713	0.646	0.615	**0.540**	**0.543**	0.512	0.466
MCD Aleatoric	0.687	0.516	0.691	0.681	0.716	0.674	0.522	0.522	0.300	0.372
MCD Epistemic	0.693	0.507	0.692	0.673	0.686	0.649	0.521	0.519	0.299	0.372
AVUQ Aleatoric	0.672	0.502	0.690	0.679	**0.718**	**0.676**	0.524	0.523	0.224	0.350
AVUQ Epistemic	**0.764**	**0.695**	**0.752**	**0.784**	0.695	0.647	0.535	0.532	**0.829**	**0.721**

Table 5: Unreliable generation detection results without thresholding.

Tables 5 and 6 report unreliable generation detection results under alternative unreliable-generation label construction choices. The no-threshold setting uses scenario provenance directly, while the mean-threshold setting uses μ rather than $\mu + 0.5\sigma$ for high-error OOD filtering. The relative trends are broadly preserved across these alternatives, suggesting that the headline results are not driven by a single task-specific threshold.

E Responsible Use Safeguards

The proposed methods are intended to support scientific imaging analysis by improving generalization and flagging potentially unreliable generations. The generated images and uncertainty scores should be used for hypothesis prioritization and model diagnostics, not as substitutes for biological or clinical validation. Any released code or model artifacts should be accompanied by documentation describing the intended research use, dataset provenance, evaluation settings, and the need for independent validation before drawing therapeutic, biological, or clinical conclusions.

F Additional Experimental Results

F.1 Controlled SFM Training and Sampling Ablation

Table 7 separates stochastic-interpolant training from the inference sampler. The deterministic FM baseline was trained from scratch with stochastic noise set to zero and without a noise head; it is not a pretrained CellFlux checkpoint. Deterministic FM and SFM use the same training data, preprocessing, U-Net trunk, optimizer, batch size, and convergence criterion. The deterministic and SFM models have 56,201,219 and 56,208,646 parameters on BBBC021, respectively, and 56,312,837 and 56,324,874 on JUMP, respectively. The noise head therefore adds 0.013% and 0.021% parameters. For the sampler comparison, we use fixed trained checkpoints, identical evaluation cohorts, integration

Method	BBBC021				JUMP					
	Unseen Pert.		Intensity Shift		Unseen Cell Lines		Unseen Plates		Intensity Shift	
	AUROC↑	AUPR↑	AUROC↑	AUPR↑	AUROC↑	AUPR↑	AUROC↑	AUPR↑	AUROC↑	AUPR↑
MAP Aleatoric	0.639	0.375	0.667	0.547	0.767	0.698	0.638	0.574	0.286	0.319
SWAG Aleatoric	0.701	0.550	0.670	0.683	0.794	0.728	0.615	0.562	0.813	0.712
SWAG Epistemic	0.676	0.523	0.678	0.684	0.782	0.714	0.627	0.574	0.879	0.797
LA Aleatoric	0.674	0.534	0.654	0.678	0.779	0.721	0.630	0.583	0.875	0.793
LA Epistemic	0.640	0.435	0.645	0.649	0.739	0.682	0.628	0.600	0.564	0.454
MCD Aleatoric	0.723	0.461	0.700	0.591	0.829	0.780	0.651	0.599	0.381	0.349
MCD Epistemic	0.727	0.453	0.706	0.595	0.777	0.728	0.634	0.579	0.371	0.346
AVUQ Aleatoric	0.710	0.467	0.699	0.602	0.841	0.796	0.666	0.614	0.302	0.324
AVUQ Epistemic	0.795	0.633	0.764	0.716	0.784	0.710	0.614	0.557	0.893	0.802

Table 6: Unreliable generation detection results with mean threshold μ.

steps, and numbers of generated images. "Naive SDE" adds inference noise without score correction; SFM with ODE sampling retains stochastic-interpolant training but omits stochastic sampling.

Table 7: Controlled training and inference comparison. Entries are FID / KID (lower is better). Bold marks the best result among these five variants for each cohort. The results use fixed trained checkpoints rather than multiple independent training runs.

Training / inference	BBBC021 Unseen Pert.	BBBC021 Intensity Shift	JUMP Unseen Cell Lines	JUMP Unseen Plates
Deterministic FM + ODE (CellFlux)	103.73 / 12.76	62.01 / 5.84	39.06 / 3.40	**17.81 / 1.04**
Deterministic FM + naive SDE	170.74 / 18.42	183.46 / 20.89	249.22 / 29.95	201.50 / 23.34
SFM + ODE	36.81 / 3.24	42.12 / 3.77	25.73 / 1.98	22.32 / 1.65
SFM + naive SDE	117.14 / 11.00	129.41 / 12.06	288.73 / 34.37	247.49 / 28.63
SFM + score-corrected SDE	**33.29 / 2.02**	**28.14 / 1.87**	**25.10 / 1.75**	18.02 / 1.23

Naive inference noise degrades both trained backbones, whereas the score-corrected SDE improves on SFM with ODE sampling in all four cohorts. The SFM ODE variant improves on the matched deterministic FM ODE baseline in three of the four cohorts.

F.2 Direct Variance Reduction with Antithetic Sampling

We compare i.i.d. and antithetic SDE samples at the same budget of $M = 4$ posterior draws and $K = 4$ SDE samples per draw. Table 8 reports antithetic-pair correlation and the variance of the estimated conditional mean across three sampling seeds. The variance-reduction factor (VRF) is the i.i.d. variance divided by the antithetic variance. Pair correlations are negative in all four cohorts, and the variance falls by 1.82–4.29 times at the same 16-sample budget.

Table 8: Equal-budget i.i.d. versus antithetic SDE sampling ($M = 4$, $K = 4$). Conditional-mean estimator variance is measured across three sampling seeds; VRF is i.i.d. variance divided by antithetic variance.

Shift	Pair corr. ρ	i.i.d. variance	Antithetic variance	VRF
BBBC021 unseen perturbations	-0.774	1.70×10^{-3}	6.44×10^{-4}	$2.64\times$
BBBC021 intensity shift	-0.773	1.70×10^{-3}	7.06×10^{-4}	$2.41\times$
JUMP unseen cell lines	-0.524	8.06×10^{-4}	4.43×10^{-4}	$1.82\times$
JUMP intensity shift	-0.748	1.40×10^{-3}	3.26×10^{-4}	$4.29\times$

F.3 Comparison with Unsupervised OOD Detectors

Table 9 compares AVUQ with ODEED, a probability-flow-ODE reconstruction detector (Le Bellier & Audebert, 2024), and PatchCore, a nominal patch-feature detector (Roth et al., 2022). The detectors were fit using only in-domain data; neither OOD examples nor unreliable-generation labels were used for fitting. The labels are used only to evaluate AUROC and AUPR. All entries in this comparison are point estimates; the ODEED and PatchCore rows do not imply the three-seed standard errors reported for methods in the main detection table. The stronger AVUQ component depends on the

shift: epistemic for both BBBC021 shifts and JUMP intensity shift, and aleatoric for JUMP unseen cell lines.

Table 9: Unsupervised detector comparison. Each entry is AUROC / AUPR (higher is better); bold marks the best point estimate in each cohort.

Method	BBBC021 Unseen Pert.	BBBC021 Intensity Shift	JUMP Unseen Cell Lines	JUMP Intensity Shift
ODEED	0.615 / 0.505	0.631 / 0.671	0.686 / 0.564	0.543 / 0.568
PatchCore	0.686 / 0.481	0.751 / 0.709	0.594 / 0.545	0.866 / 0.753
AVUQ Aleatoric	0.742 / 0.523	0.729 / 0.663	**0.869 / 0.835**	0.309 / 0.335
AVUQ Epistemic	**0.809 / 0.698**	**0.785 / 0.768**	0.794 / 0.726	**0.890 / 0.799**

F.4 JUMP Unseen-Plate Detection

JUMP unseen plates is a relatively mild shift for SFM and yields fewer unreliable generations, so the main detection table focuses on the four more failure-rich cohorts. Table 10 reports the available unseen-plate detection point estimates. AVUQ Aleatoric has the highest AUROC, while LA Aleatoric has the highest AUPR.

Table 10: JUMP unseen-plate unreliable-generation detection. Point estimates are AUROC and AUPR (higher is better).

Method	AUROC	AUPR
MAP Aleatoric	0.689	0.607
MCD-DFM Epistemic	0.571	0.509
SWAG Aleatoric	0.728	0.641
SWAG Epistemic	0.727	0.657
LA Aleatoric	0.732	**0.670**
LA Epistemic	0.697	0.629
MCD Aleatoric	0.734	0.657
MCD Epistemic	0.719	0.647
AVUQ Aleatoric	**0.737**	0.661
AVUQ Epistemic	0.665	0.580

F.5 Mechanism-of-Action Classification of Generated Cell Images

Table 11 evaluates top-1 mechanism-of-action (MoA) accuracy using a pretrained BBBC021 classifier (13 classes; 7.69% uniform chance). Its 95% intervals use 2,000 perturbation-stratified bootstrap resamples. The columns separate intensity shift from its combination with unseen perturbations; real target images give a reference. SFM leads the generators in both cohorts, but its interval overlaps the closest baselines only under the combined shift.

Table 11: Top-1 MoA accuracy (95% bootstrap confidence interval). Higher is better. The same pretrained classifier is evaluated on generated and real target images.

Method	Intensity shift	Intensity shift + unseen perturbations
Uniform chance (13 classes)	7.69%	7.69%
BBDM	7.32% (6.49–8.15%)	9.93% (8.33–11.61%)
GOUB	20.13% (18.90–21.36%)	3.20% (2.24–4.16%)
UniDB	25.62% (24.36–26.97%)	5.68% (4.40–7.05%)
UNSB	4.52% (3.92–5.20%)	10.61% (9.01–12.45%)
SDEdit	43.38% (42.07–44.64%)	5.28% (4.24–6.41%)
CellFlux	36.43% (35.14–37.63%)	10.13% (8.53–11.73%)
SFM	**48.47% (46.93–49.96%)**	**12.89% (11.13–14.73%)**
Real target images	66.37% (64.97–67.72%)	43.23% (40.59–45.96%)

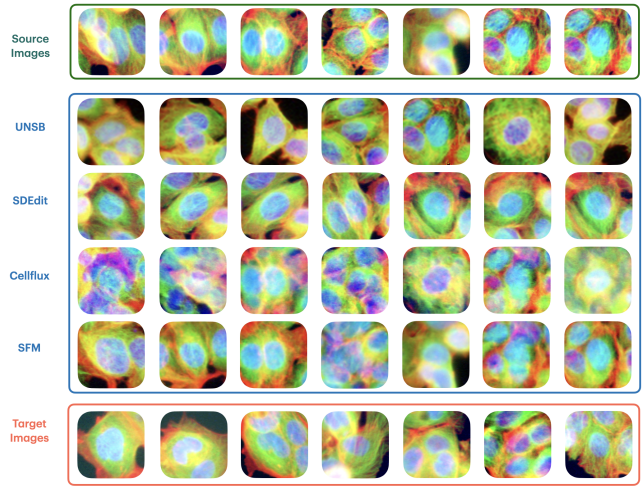


Figure 4: Examples of generated images from different methods on BBBC021 under the Unseen Pert. OOD scenario, compared with source images and ground-truth target images.

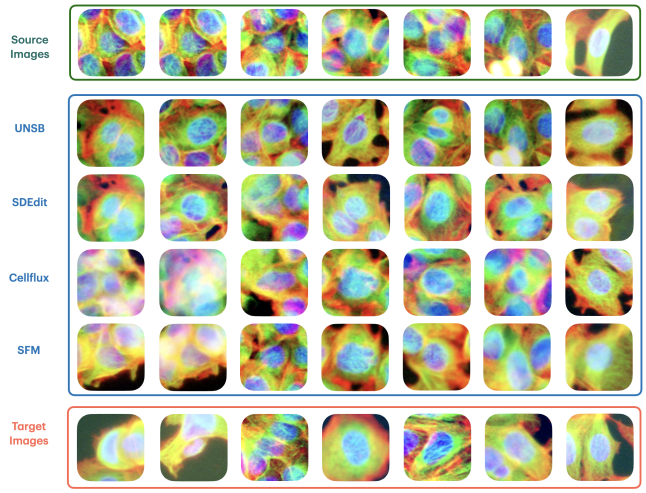


Figure 5: Examples of generated images from different methods on BBBC021 under the Intensity Shift OOD scenario, compared with source images and ground-truth target images.

F.6 Visualizing SFM Generalization

Figures 4, 5, 6, and 7 show additional examples of generated images from different methods on BBBC021 and JUMP under various OOD scenarios, compared with the source images and ground-truth target images.

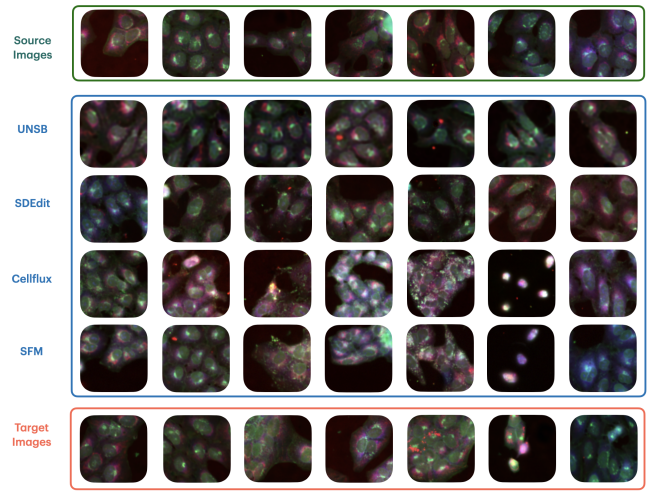


Figure 6: Examples of generated images from different methods on JUMP under the Unseen Cell Lines OOD scenario, compared with source images and ground-truth target images.

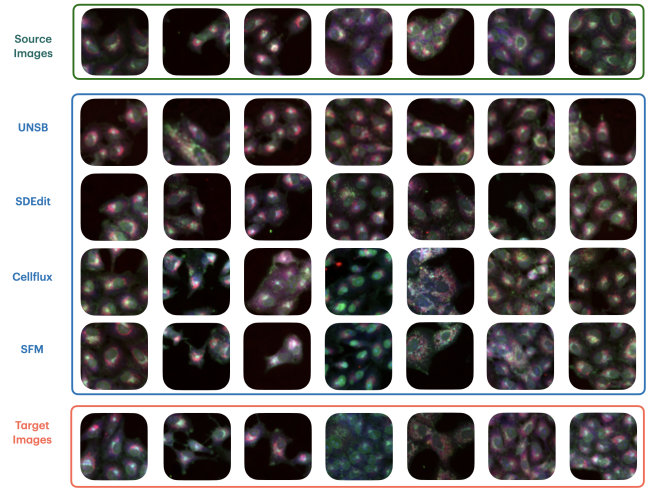


Figure 7: Examples of generated images from different methods on JUMP under the Unseen Plates OOD scenario, compared with source images and ground-truth target images.