Stochastic Perturbations Improve Distribution-to-Distribution Generative Models

Supplementary Material

A. Inference algorithm

Algorithm 2 Inference

Require: Source sample x_0, learned fields v_θ, η_ϕ, interpolant noise schedule γ_t, inference source noise scale ϵ, diffusion coefficient σ_t, step size Δt, numerical solver function `Step`

$z \sim \mathcal{N}(0, I)$
$\tilde{x}_0 = x_0 + \epsilon z$
Initialize $x_t = \tilde{x}_0$ at $t = 0$
for $t = 0$ to 1 **do**
    $x_{t+\Delta t} = \texttt{Step}\big(x_t, t, v_\theta, \eta_\theta, \gamma_t, \sigma_t\big)$
end for

B. Theorems and derivations

To formally state the theorem, we need to first define the population risk and empirical risk for flow matching and our loss. Generally, both losses can be written as

$$\mathcal{L}_{\text{FM}}(\theta, t) = \mathbb{E}_{x_t \sim p_t(x_t)} \left[\frac{1}{2} \| v_t^\theta(x_t) - v_t^{\text{FM}}(x_t) \|^2 \right] \quad (14)$$

$$\mathcal{L}_{\text{SI}}(\theta, t) = \mathbb{E}_{x_t \sim q_t(x_t)} \left[\frac{1}{2} \| v_t^\theta(x_t) - v_t^{\text{SI}}(x_t) \|^2 \right] \quad (15)$$

where $v_t^*(x_t)$ is the marginal velocity at x_t, and $p_t(x_t), q_t(x_t)$ are the ground-truth distributions to draw x_t for each loss.

Theorem 1 (Informal)**.** *Let $\mathcal{L}_{FM}(\theta, t), \mathcal{L}_{SI}(\theta, t)$ be the population risk of flow matching and our stochastic injection loss at time t, and $\hat{\mathcal{L}}_{FM}(\theta, t), \hat{\mathcal{L}}_{SI}(\theta, t)$ be their empirical risks with n i.i.d. samples. Let $p_t(x_t), q_t(x_t)$ be the respective population distribution of x_t, and $\hat{p}_t(x_t), \hat{q}_t(x_t)$ be their empirical distributions, the 1-Wasserstein distance $\mathbb{W}_1(p_t, \hat{p}_t), \mathbb{W}_1(q_t, \hat{q}_t)$ characterizes each loss' generalization gap. Moreover, $\mathbb{W}_1(q_t, \hat{q}_t) \leq \mathbb{W}_1(p_t, \hat{p}_t)$.*

Let $\mathcal{L}_{\text{FM}}(\theta, t), \mathcal{L}_{\text{SI}}(\theta, t)$ be the population risk of Flow Matching and our stochastic injection loss at $t \in [0, 1]$, and $\hat{\mathcal{L}}_{\text{FM}}(\theta, t), \hat{\mathcal{L}}_{\text{SI}}(\theta, t)$ be their empirical risks with n i.i.d. samples. Let $p_t(x_t), q_t(x_t)$ be the respective population distribution of x_t, and $\hat{p}_t(x_t), \hat{q}_t(x_t)$ be their empirical distributions, and assume each loss is Lipschitz continuous with regard to x_t, we have

$$|\mathcal{L}_{\text{FM}}(\theta, t) - \hat{\mathcal{L}}_{\text{FM}}(\theta, t)| \leq L \mathbb{W}_1(p_t, \hat{p}_t), \quad (16)$$

$$|\mathcal{L}_{\text{SI}}(\theta, t) - \hat{\mathcal{L}}_{\text{SI}}(\theta, t)| \leq L \mathbb{W}_1(q_t, \hat{q}_t) \quad (17)$$

for some constant L. Moreover, $\mathbb{W}_1(q_t, \hat{q}_t) \leq \mathbb{W}_1(p_t, \hat{p}_t)$.

Proof. Without loss of generality we let $\mathcal{G}_\theta^{\text{FM}}(x_t) = \frac{1}{2} \| v_t^\theta(x_t) - v_t^{\text{FM}}(x_t) \|^2$ with Lipschitz constant L_{FM}, and similarly $\mathcal{G}_\theta^{\text{SI}}(x_t)$ has Lipschitz constant L_{SI}, and so

$$\mathcal{L}_{\text{FM}}(\theta, t) = \mathbb{E}_{x_t \sim p_t(x_t)} \left[\mathcal{G}_\theta(x_t) \right], \quad (18)$$

$$\mathcal{L}_{\text{SI}}(\theta, t) = \mathbb{E}_{x_t \sim q_t(x_t)} \left[\mathcal{G}_\theta(x_t) \right] \quad (19)$$

By Kantorovich–Rubinstein duality of $\mathbb{W}_1(p_t, \hat{p}_t)$, we have

$$\left| \mathcal{L}_{\text{FM}}(\theta, t) - \hat{\mathcal{L}}_{\text{FM}}(\theta, t) \right| \quad (20)$$

$$= \left| \mathbb{E}_{x_t \sim p_t(x_t)} \left[\mathcal{G}_\theta(x_t) \right] - \mathbb{E}_{x_t \sim \hat{p}_t(x_t)} \left[\mathcal{G}_\theta(x_t) \right] \right| \quad (21)$$

$$\leq L_{\text{FM}} \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{x_t \sim p_t(x_t)} \left[f(x_t) \right] - \mathbb{E}_{x_t \sim \hat{p}_t(x_t)} \left[f(x_t) \right] \right| \quad (22)$$

$$\leq \max\{L_{\text{FM}}, L_{\text{SI}}\} \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{x_t \sim p_t(x_t)} \left[f(x_t) \right] - \right. \quad (23)$$

$$\left. \mathbb{E}_{x_t \sim \hat{p}_t(x_t)} \left[f(x_t) \right] \right| \quad (24)$$

$$= L \mathbb{W}_1(p_t, \hat{p}_t) \quad (25)$$

where we let $L = \max\{L_{\text{FM}}, L_{\text{SI}}\}$ and $\mathcal{F}$ is a function set with Lipschitz constant of at most 1. The same conclusion holds for $\mathcal{L}_{\text{SI}}(\theta, t)$.

For the final result, since $q_t(x_t)$ is determinstic interpolation (drawn from $p_t(x_t)$) with Gaussian noise, it can be equivalently written as $q_t = p_t * N_t$ where $N_t(x) = \mathcal{N}(0, \sigma_t^2 I)$, a Gaussian convolution of $p_t(x_t)$ with variance σ_t^2. Similarly $\hat{q}_t = \hat{p}_t * N_t$. Therefore,

$$\mathbb{W}_1(q_t, \hat{q}_t) = \mathbb{W}_1(p_t * N_t, \hat{p}_t * N_t) \overset{(i)}{\leq} \mathbb{W}_1(p_t, \hat{p}_t) \quad (26)$$

where (i) is due to Wasserstein-reducing property of Gaussian smoothing [8]. □

We remark that $\mathbb{W}_1(p_t, \hat{p}_t)$ is a measure of an upper bound of the generalization gap, and it does not strictly characterize the gap, so the exact relationship between the two generalization gaps cannot be measured precisely. However, we use the 1-Wasserstein distance as an approximation of the gap to give a rough intuition on why injecting stochastic noise can help test performance.

Lemma 1. *If the source distribution $p_0(x_0)$ is a mixture of delta distributions $\frac{1}{n} \sum_{i=0}^{n} \delta(x - x_i)$ with sample size n, then the ground-truth probability-flow ODE can only recover a mixture of delta distributions with sample size n.*

Proof. We know that the ground-truth probability-flow ODE path does not cross, and therefore the ground-truth ODE flow is a one-to-one function. Let $\Phi(x_0)$ denote the ground-truth flow path following probability-flow ODE, and consider the pushforward distribution via the flow path as $(\Phi \# p_0)(x_1) = \frac{1}{n}\sum_{i=0}^{n} \Phi \# \delta(x_0 - x_i) = \frac{1}{n}\sum_{i=0}^{n} \delta(x_1 - \Phi(x_i))$, which is a mixture of delta distributions with sample size n. □

C. Datasets

We describe each dataset below.

- BBBC. We use the BBBC021v1 image set [7] from the Broad Bioimage Benchmark Collection [24]. This dataset contains fluorescent microscopy of cells treated with 26 small molecule chemicals, forming a conditional target distribution for each chemical. Three color channels correspond to DNA, F-actin, and beta-tubulin markers.
- SEASONET. This dataset contains multi-spectral aerial image patches covering the surface of Germany from the Sentinel-2 mission, collected from April 2018 to February 2019 [19]. The images are available in standard RGB channels and sorted into each of four seasons. We use only the summer and winter splits for the source and target distributions respectively.
- YOSEMITE. We use images of Yosemite National Park collected by Zhu et al. [38] via the Flickr API. The dataset separates images taken in the summer and images taken in the winter, which we use as the source and target distributions.
- MIMIC-CXR. MIMIC-CXR is a medical imaging dataset of chest radiographs [16]. We filter to those images with the antero-posterior view angle. Pleural effusion is a condition characterized by fluid around the lungs. The source distribution is defined as scans from patients with pleural effusion value of 0.0, and the target distribution scans from patients with pleural effusion value of 1.0. The scans are in single-channel grayscale. We resize them to 256×256.
- GALAXIESML. We use galaxy images from the Hyper-Suprime-Cam (HSC) Survey [1] as processed by Do et al. [11]. This dataset contains five photometric bands (g, r, i, z, y) as well as spectroscopically confirmed redshifts. We use the g, r, and i channels to construct a 3-channel image. The images with redshift values 0.3-0.5 are used as the source distribution and images obtained at redshift values 0.5-0.7 are used as the target distribution.

Datasets statistics are summarized in Table 4. All FID numbers are reported over the held out test sets.

D. Training details

D.1. Experiments on CONCENTRICSHELLS

In this task, the source distribution (inner shell) is a hypersphere centered at the origin with radius 1 and the target distribution (outer shell) is a hypersphere centered at the origin with radius 2. For both the source and target distributions, each sample is obtained by sampling over the d-dimensional shell, then perturbing this with a random normal noise component with standard deviation 0.1.

For the data dimension scaling experiment, each training run operates over 1024 samples from the source distribution (inner shell) and 1024 samples from the target distribution (outer shell). For the dataset size scaling experiment, the data dimension is fixed to 512. We use the Adam optimizer at learning rate 0.01 and batch size 256. The velocity field is fit by a simple 4-layer MLP with hidden dimension 64 and an ELU non-linearity between each fully connected layer.

All metrics are reported over a test set of 512 samples. We compute the Sinkhorn distance with entropy regularization of 0.1.

D.2. Variational autoencoder

We learn velocity and score fields in the latent space of a VAE for each of the image datasets introduced in Section 4.1. During training and inference, we use different variational autoencoders that is best adapted to each dataset. They are first trained from the same data as what is available for the distribution learning, and the VAE weights are subsequently frozen. We same architecture as the $f = 4$ autoencoder from Rombach et al. [28]. The VAE for BBBC is trained from scratch. The VAE for each of SEASONET, MIMIC-CXR, and GALAXIESML are fine-tuned from Rombach et al. [28]'s kl-f4 checkpoint, trained for 176991 steps, at the default KL regularization penalty of 1×10^{-6}. For YOSEMITE, we directly use their pre-trained autoencoder as we found fine-tuning on YOSEMITE led to overfitting and performance degradation, likely due to the small dataset size.

D.3. Main experiments

On all datasets, models were trained with constant learning rate 1×10^{-4} with the AdamW optimizer (betas 0.9 and 0.95). We maintain an EMA-weighted copy of the model with 0.999 decay. For the conditional dataset BBBC, we drop class labels with probability 0.2. Each epoch iterates over all training data in the target distribution while randomly sampling training data in the source distribution. We train for 200, 100, 2000, 60, 80 epochs for each of the datasets BBBC, SEASONET, YOSEMITE, MIMIC-CXR, and GALAXIESML respectively, based on observed convergence. When training with the two-stage scheme, some fraction of these epochs are reserved for the noise-to-target

Table 4. Dataset statistics.

	# train(A)	# train(B)	# test(A)	# test(B)	resolution	domain
BBBC	63,781	6,210	690	7,119	256×256	cell microscopy
SeasoNet	235,826	104,432	1,024	1,024	120×120	satellite
Yosemite	1,231	962	309	238	256×256	natural
MIMIC-CXR	16,038	44,372	1,024	1,024	256×256	medical x-ray
GalaxiesML	35,725	45,741	1,024	1,024	127×127	astronomy

stage. Hence the two-stage training does not incur additional compute.

Sampling is performed with the Heun solver with 50 inference steps (corresponding to 100 NFEs) unless stated otherwise. The stochastic variant of the Heun solver [17] was used for the ODE/SDE comparison experiments (Figure 5c). Similar to some prior work [25], we chose the diffusion coefficient $\sigma_t^2/2 = \sin^2(\pi t)$. We also experimented with a time-independent σ_t but found it performed worse than a schedule that is tapered at the $t = 0$ and $t = 1$ endpoints. Additionally, we set the diffusion coefficient to 0 within a margin $\epsilon = 1 \times 10^{-3}$ near the endpoints, to avoid the numerical instability caused by the factor of γ^{-1}. We find this is crucial to obtain reasonable samples.

For the DDIB and SDEdit baselines, we require access to a generative model that can conditionally flow from noise to both the source and the target. For full comparability, we train a noise-to-source/target flow matching model for each dataset, keeping all hyperparameters consistent with the flow baseline where applicable.

E. Evaluation metrics in domain-specific embedding Spaces

FID [14] is computed in Inception-v3 feature space, which is a suboptimal representation space for specialized scientific images. To address this, we additionally compute Fréchet Distance (FD) and Kernel Distance (KD) [6] in domain-specific embedding spaces, providing convergent evidence that our improvements reflect genuine gains in domain-relevant fidelity rather than alignment with ImageNet statistics.

We use the following encoders for each scientific domain dataset:

- BBBC (cell microscopy): The mode-of-action (MoA) classifier from CellFlux [37], trained on the same BBBC021 dataset, whose representations capture the morphological signatures that distinguish different chemical interventions.
- SeasoNet (satellite imagery): RemoteCLIP [21], a vision-language foundation model contrastively pretrained on a large collection of remote sensing datasets, providing rich semantic representations of satellite imagery.
- MIMIC-CXR (chest radiographs): MedSigLIP [30], an encoder from Google's Health AI Developer Foundations program, adapted from SigLIP [36] on diverse medical image-text pairs including MIMIC-CXR itself.
- GalaxiesML (galaxy images): Zoobot [34], a foundation model for galaxy morphology pretrained on millions of Galaxy Zoo volunteer annotations spanning multiple telescope surveys.

Yosemite consists of natural photographs for which the Inception-v3 encoder is already well-suited.

Table 5 shows that our stochastic injections achieve the lowest FD and KD, reflecting the same trends as our main FID results and reinforcing that the quality improvements generalize across embedding spaces.

F. Hyperparameter sensitivity

We investigate the sensitivity of our results to noise injection hyperparameters. Our main results use a sine-squared interpolant noise schedule at scale 1.0; this is tuned only on BBBC and applied across all other datasets without further adjustment. While dataset-specific tuning could yield further gains, we argue that our defaults constitute a broadly robust configuration.

To verify this, we run additional ablations in which we train from scratch with a single hyperparameter perturbed at a time, keeping all others at their default values. We choose BBBC and GalaxiesML for this sensitivity study.

Table 6 reports FID across low, mid, and high settings for the three noise hyperparameters: the proportion of epochs spent in the noise-to-target pretraining stage (Section 3.1), the scale of Gaussian noise perturbing the source samples (Section 3.2), and the scale of noise perturbing the interpolant (Section 3.3). The low and high settings differ from the default mid setting by approximately 20-25%. Across both datasets and all three hyperparameters, all reasonable choices achieve strong gains over the baselines (c.f. Table 1). Dataset-specific tuning occasionally yields marginal further improvement, but the differences between the low, mid, and high settings are modest compared to the performance gains from our stochastic injections.

Table 5. Fréchet $(\cdot)_F$ and Kernel $(\cdot)_K$ distances with domain-specific embeddings. $^*\times 10^{-2}$; $^\dagger\times 10^{-2}$; $^\ddagger\times 10^{7}$.

	Fréchet Distance (FD)				Kernel Distance (KD)			
	BBBC	SeasoNet*	MIMIC	Galaxies	BBBC	SeasoNet†	MIMIC	Galaxies‡
UNSB	212.7	23.3	7.8	5.2e4	843.3	48.4	15.0	2.5e3
DDIB	22.5	12.3	2.9	116.0	49.8	21.9	3.5	5.1
SDEdit	237.7	62.0	25.9	5.9e4	1.1e3	138.4	46.5	8.7e3
Flow	31.8	9.9	2.6	214.9	74.2	17.6	3.5	4.6
Flow w. stochastic (ours)	**13.6**	**8.3**	**2.1**	**89.4**	**15.5**	**14.3**	**2.0**	**3.4**

Table 6. Hyperparameter sensitivity ablations on BBBC and GalaxiesML (FID, ↓). We perturb one hyperparameter at a time, keeping the others at their default (mid) values. *Epochs*: proportion of training spent in the noise-to-target pretraining stage; low/mid/high corresponds to 80/100/120 pretraining epochs out of 200 total for BBBC, and 30/40/50 out of 80 total for GalaxiesML. *Source noise scale*: standard deviation of Gaussian noise added to source samples; low/mid/high = 0.75/1.0/1.25. *Interpolant noise scale*: scale parameter a of the sine-squared noise schedule $\gamma_t = a\sin^2(\pi t)$; low/mid/high = 0.75/1.0/1.25.

	BBBC			GalaxiesML		
	Low	Mid	High	Low	Mid	High
Epochs in noise-to-target	20.3	19.9	19.6	7.4	7.6	7.9
Source noise scale	20.5	19.9	19.6	8.2	7.4	7.8
Interpolant noise scale	20.7	19.9	20.2	7.7	7.4	7.5

G. Further qualitative examples

Figure 6 shows a sample generation from each baseline and from our proposed solution of flow enhanced with stochastic injections. Qualitative observations support the results of Table 1. Our proposed method obtains the highest quality visual samples, which both resemble other images in the target distribution while retaining the visual correspondence with the source. Of the baselines, DDIB is the strongest, but still can miss subtle details in distribution matching, for example in MIMIC-CXR showing limited evidence of fluid in the lung cavity. Samples from UNSB tend to be highly similar to the source image, suggesting its solution collapses to close to the identity map and fails to capture some subtleties of the true transformation. In the case of GALAXIESML, UNSB learns some spurious artefacts visible as the haziness in the center of the example. SDEdit often produces samples that appear noise corrupted. Finally, our stochastic injections also improve over standard flow matching. For instance, it resolves the "overfitting to snow" effect in the sample for YOSEMITE, and better maintains the boundaries of the forested land in the sample for SEASONET.

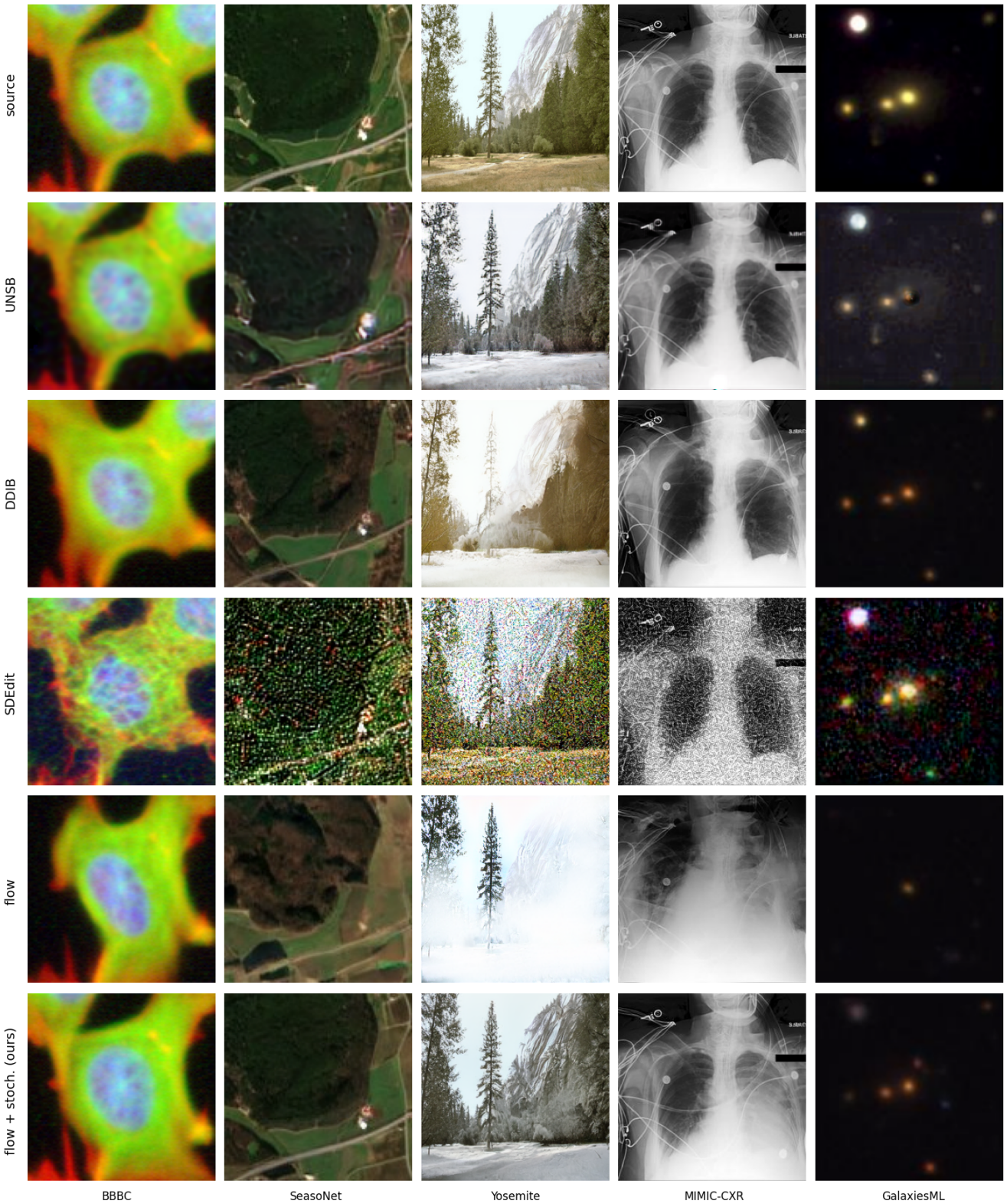

Figure 6. Qualitative examples for each method from Table 1.