∞-Brush: Controllable Large Image Synthesis with Diffusion Models in Infinite Dimensions —Supplementary Material—

Minh-Quan Le*, Alexandros Graikos*, Srikar Yellapragada, Rajarsi Gupta, Joel Saltz, and Dimitris Samaras

Stony Brook University
{mile,agraikos,myellapragad,samaras}@cs.stonybrook.edu

In this supplementary material, we present a detailed derivation of our proposed method ∞-Brush and extensive experimental results. In section A, we provide the details of forward process, reverse process, full proof of Proposition 1 and Theorem 1. In section B, we perform experiments to compare the long-range dependency between the patch-based method [3] and ∞-Brush, zero-shot classification using a Vision-Language Model (VLM) Quilt, application of synthetic data on downstream task, and ablation study on % pixels for training. We further demonstrate the qualitative results of our method on $4096 \times 4096, 2048 \times 2048,$ and 1024×1024 resolutions of TCGA-BRCA [1] and NAIP [6] datasets along with failure cases generated with our method.

A Conditional Diffusion Models in Function Space

Forward process. The forward process of a conditional diffusion model in function space is defined as a discrete-time Markov chain that incrementally perturbs probability measure $\mathbb{Q}_{\text{data}}$ towards a Gaussian measure $\mathcal{N}(\mathbf{m}, \mathbf{C})$ with a zero mean and a specified covariance operator $\mathbf{C}$. It is a time-indexed process where each step $\mathbf{u}_t$ is obtained by applying a transformation to the previous step $\mathbf{u}_{t-1}$, which involves a scaling factor $\sqrt{1-\beta_t}\mathbf{u}_{t-1}$ related to the variance schedule β, and adding scaled Gaussian noise $\sqrt{\beta_t}\boldsymbol{\xi}_t$ with $\boldsymbol{\xi}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$:

$$\mathbf{u}_t = \sqrt{1-\beta_t}\mathbf{u}_{t-1} + \sqrt{\beta_t}\boldsymbol{\xi}_t \qquad t = 1, 2, \ldots, T. \tag{1}$$

Similar to diffusion models in finite dimensions, the forward process in function space also admits sampling $\mathbf{u}_t$ at an arbitrary timestep t in closed form. For $\bar{\alpha}_t = \prod_{i=1}^{t}(1-\beta_t)$, we have:

$$\begin{aligned}
\mathbf{u}_t &= \sqrt{1-\beta_t}\mathbf{u}_{t-1} + \sqrt{\beta_t}\boldsymbol{\xi}_t \quad ;\text{where } \boldsymbol{\xi}_t, \boldsymbol{\xi}_{t-1}, \cdots \sim \mathcal{N}(\mathbf{0}, \mathbf{C}) \\
&= \sqrt{(1-\beta_t)(1-\beta_{t-1})}\mathbf{u}_{t-2} + \sqrt{(1-\beta_t)\beta_{t-1}}\boldsymbol{\xi}_{t-1} + \sqrt{\beta_t}\boldsymbol{\xi}_t \\
&= \sqrt{(1-\beta_t)(1-\beta_{t-1})}\mathbf{u}_{t-2} + \sqrt{1-\alpha_t\alpha_{t-1}}\bar{\boldsymbol{\xi}}_{t-1} \\
&= \ldots \\
&= \sqrt{\bar{\alpha}_t}\mathbf{u}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\xi}
\end{aligned} \tag{2}$$

* Equal contribution

Based on the above analysis, we obtain:

$$\mathbb{Q}\left(\mathbf{u}_t|\mathbf{u}_0\right) = \mathcal{N}\left(\mathbf{u}_t; \sqrt{\bar{\alpha}_t}\mathbf{u}_0, (1-\bar{\alpha}_t)\mathbf{C}\right). \tag{3}$$

In the context of image generation, we discretize the function $\mathbf{u}_j$ on the mesh $\mathbf{x}_j = \{\mathbf{x}_j^{(i)}\}_{1\leq i\leq N} \subset \mathcal{X}$ by sampling N coordinates of each image, which results in a non-smooth input space. To achieve a smoother function representation, a smoothing operator [4, 5] $\mathbf{A} : \mathcal{H} \to \mathcal{H}$, *e.g.* a truncated Gaussian kernel, is applied to approximate the rough inputs within the function space $\mathcal{H}$:

$$\mathbb{Q}\left(\mathbf{u}_t|\mathbf{u}_0\right) = \mathcal{N}\left(\mathbf{u}_t; \sqrt{\bar{\alpha}_t}\mathbf{A}\mathbf{u}_0, (1-\bar{\alpha}_t)\mathbf{A}\mathbf{C}\mathbf{A}^T\right). \tag{4}$$

Reverse process. The reverse process in the diffusion model framework is achieved by iteratively denoising from the Gaussian measure $\mathcal{N}(\mathbf{m}, \mathbf{C})$ back towards the probability measure $\mathbb{Q}_0 = \mathbb{Q}_{\text{data}}$. We use a variational approach to approximate posterior measures with a variational family of measures on $\mathcal{H}$ and incorporate the conditional embedding $\mathbf{e}$ to control the generation process. We model the underlying posterior measure $\mathbb{Q}(\mathbf{u}_{t-1}|\mathbf{u}_t)$ with a conditional Gaussian measure:

$$\mathbb{P}_\theta(\mathbf{u}_{t-1}|\mathbf{u}_t, \mathbf{e}) = \mathcal{N}\left(\mathbf{u}_{t-1}; \mathbf{m}_\theta(\mathbf{u}_t, \mathbf{e}, t), \mathbf{A}\mathbf{C}_\theta(\mathbf{u}_t, \mathbf{e}, t)\mathbf{A}^T\right). \tag{5}$$

Likewise, we are able to derive a closed-form representation of the forward process posteriors, which are tractable when conditioned on $\mathbf{u}_0$:

$$\mathbb{Q}\left(\mathbf{u}_{t-1}|\mathbf{u}_t, \mathbf{u}_0\right) = \mathcal{N}\left(\mathbf{u}_{t-1}; \tilde{\mathbf{m}}_t(\mathbf{u}_t, \mathbf{u}_0), \tilde{\beta}_t\mathbf{C}\right). \tag{6}$$

Using Bayes' rule, we obtain:

$$\begin{aligned}
&\mathbb{Q}(\mathbf{u}_{t-1}|\mathbf{u}_t, \mathbf{u}_0) = \mathbb{Q}(\mathbf{u}_t|\mathbf{u}_{t-1}, \mathbf{u}_0)\frac{\mathbb{Q}(\mathbf{u}_{t-1}|\mathbf{u}_0)}{\mathbb{Q}(\mathbf{u}_t|\mathbf{u}_0)} \\
&\propto \exp\Big(-\frac{1}{2}\big(\frac{\langle \mathbf{C}^{-1}(\mathbf{u}_t - \sqrt{\alpha_t}\mathbf{u}_{t-1}), \mathbf{u}_t - \sqrt{\alpha_t}\mathbf{u}_{t-1}\rangle}{\beta_t} + \frac{\langle \mathbf{C}^{-1}(\mathbf{u}_{t-1} - \sqrt{\bar{\alpha}_{t-1}}\mathbf{A}\mathbf{u}_0), \mathbf{u}_{t-1} - \sqrt{\bar{\alpha}_{t-1}}\mathbf{A}\mathbf{u}_0\rangle}{1-\bar{\alpha}_{t-1}} \\
&- \frac{\langle \mathbf{C}^{-1}(\mathbf{u}_t - \sqrt{\bar{\alpha}_t}\mathbf{A}\mathbf{u}_0), \mathbf{u}_t - \sqrt{\bar{\alpha}_t}\mathbf{A}\mathbf{u}_0\rangle}{1-\bar{\alpha}_t}\big)\Big) \\
&= \exp\Big(-\frac{1}{2}\big((\frac{\alpha_t}{\beta_t} + \frac{1}{1-\bar{\alpha}_{t-1}})\langle \mathbf{C}^{-1}\mathbf{u}_{t-1}, \mathbf{u}_{t-1}\rangle - 2\langle \mathbf{C}^{-1}(\frac{\sqrt{\alpha_t}}{\beta_t}\mathbf{u}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}}\mathbf{A}\mathbf{u}_0), \mathbf{u}_{t-1}\rangle + C(\mathbf{u}_t, \mathbf{u}_0)\big)\Big),
\end{aligned} \tag{7}$$

where $C(\mathbf{u}_t, \mathbf{u}_0)$ is some function not involving $\mathbf{u}_{t-1}$ and details are omitted. Following the standard Gaussian density function, the mean and covariance of $\mathbb{Q}(\mathbf{u}_{t-1}|\mathbf{u}_t, \mathbf{u}_0)$ can be parameterized as follows (recall that $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^T \alpha_i$):

$$\begin{aligned}
\tilde{\mathbf{m}}_t(\mathbf{u}_t, \mathbf{u}_0) &= (\frac{\sqrt{\alpha_t}}{\beta_t}\mathbf{u}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}}\mathbf{A}\mathbf{u}_0)/(\frac{\alpha_t}{\beta_t} + \frac{1}{1-\bar{\alpha}_{t-1}}) \\
&= (\frac{\sqrt{\alpha_t}}{\beta_t}\mathbf{u}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}}\mathbf{A}\mathbf{u}_0)\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\cdot\beta_t \\
&= \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{A}\mathbf{u}_0 + \frac{\sqrt{1-\beta_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{u}_t.
\end{aligned} \tag{8}$$

$$\tilde{\beta}_t = 1/(\frac{\alpha_t}{\beta_t} + \frac{1}{1-\bar{\alpha}_{t-1}}) = 1/(\frac{\alpha_t - \bar{\alpha}_t + \beta_t}{\beta_t(1-\bar{\alpha}_{t-1})}) = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \cdot \beta_t. \tag{9}$$

Proposition 1 (Learning Objective). *The cross-entropy of conditional diffusion models in function space has a variational upper bound of*

$$\mathcal{L}_{\text{CE}} = -\mathbb{E}_{\mathbb{Q}} \log \mathbb{P}_\theta(\mathbf{u}_0|\mathbf{e}) \leq \mathbb{E}_{\mathbb{Q}} \Bigg[\underbrace{\text{KL}(\mathbb{Q}(\mathbf{u}_T|\mathbf{u}_0) \parallel \mathbb{P}_\theta(\mathbf{u}_T))}_{\mathcal{L}_T} \underbrace{- \log \mathbb{P}_\theta(\mathbf{u}_0|\mathbf{u}_1, \mathbf{e})}_{\mathcal{L}_0} + \sum_{t=2}^{T} \underbrace{\text{KL}(\mathbb{Q}(\mathbf{u}_{t-1}|\mathbf{u}_t, \mathbf{u}_0) \parallel \mathbb{P}_\theta(\mathbf{u}_{t-1}|\mathbf{u}_t, \mathbf{e})}_{\mathcal{L}_{t-1}} \Bigg]. \tag{10}$$

Proof. The conditional diffusion model in function space is trained to minimize the cross entropy as the learning objective, which is equivalent to minimize variational upper bound (VUB):

$$\begin{aligned}
L_{\text{CE}} &= -\mathbb{E}_{\mathbb{Q}(\mathbf{u}_0|\mathbf{e})} \log \mathbb{P}_\theta(\mathbf{u}_0|\mathbf{e}) \\
&= -\mathbb{E}_{\mathbb{Q}(\mathbf{u}_0|\mathbf{e})} \log \Big(\int \mathbb{P}_\theta(\mathbf{u}_{0:T}|\mathbf{e}) d\mathbf{u}_{1:T} \Big) \\
&= -\mathbb{E}_{\mathbb{Q}(\mathbf{u}_0|\mathbf{e})} \log \Big(\int \mathbb{Q}(\mathbf{u}_{1:T}|\mathbf{u}_0, \mathbf{e}) \frac{\mathbb{P}_\theta(\mathbf{u}_{0:T}|\mathbf{e})}{\mathbb{Q}(\mathbf{u}_{1:T}|\mathbf{u}_0, \mathbf{e})} d\mathbf{u}_{1:T} \Big) \\
&= -\mathbb{E}_{\mathbb{Q}(\mathbf{u}_0|\mathbf{e})} \log \Big(\mathbb{E}_{\mathbb{Q}(\mathbf{u}_{1:T}|\mathbf{u}_0, \mathbf{e})} \frac{\mathbb{P}_\theta(\mathbf{u}_{0:T}|\mathbf{e})}{\mathbb{Q}(\mathbf{u}_{1:T}|\mathbf{u}_0, \mathbf{e})} \Big) \\
&\leq -\mathbb{E}_{\mathbb{Q}(\mathbf{u}_{0:T}|\mathbf{e})} \log \frac{\mathbb{P}_\theta(\mathbf{u}_{0:T}|\mathbf{e})}{\mathbb{Q}(\mathbf{u}_{1:T}|\mathbf{u}_0, \mathbf{e})} \\
&= \mathbb{E}_{\mathbb{Q}(\mathbf{u}_{0:T}|\mathbf{e})} \Big[\log \frac{\mathbb{Q}(\mathbf{u}_{1:T}|\mathbf{u}_0, \mathbf{e})}{\mathbb{P}_\theta(\mathbf{u}_{0:T}|\mathbf{e})} \Big] \\
&= \mathbb{E}_{\mathbb{Q}(\mathbf{u}_{0:T}|\mathbf{e})} \Big[\log \frac{\mathbb{Q}(\mathbf{u}_{1:T}|\mathbf{u}_0)}{\mathbb{P}_\theta(\mathbf{u}_{0:T}|\mathbf{e})} \Big] = L_{\text{VUB}}.
\end{aligned} \tag{11}$$

To convert each term in the equation to be analytically computable, the objective can be further rewritten to be a combination of several KL-divergence

and entropy terms:

$$
\begin{aligned}
L_{\text{VUB}} &= \mathbb{E}_{\mathbb{Q}(\mathbf{u}_{0:T}|\mathbf{e})}\Big[\log \frac{\mathbb{Q}(\mathbf{u}_{1:T}|\mathbf{u}_0)}{\mathbb{P}_\theta(\mathbf{u}_{0:T}|\mathbf{e})}\Big] \\
&= \mathbb{E}_{\mathbb{Q}}\Big[\log \frac{\prod_{t=1}^{T}\mathbb{Q}(\mathbf{u}_t|\mathbf{u}_{t-1})}{\mathbb{P}_\theta(\mathbf{u}_T)\prod_{t=1}^{T}\mathbb{P}_\theta(\mathbf{u}_{t-1}|\mathbf{u}_t,\mathbf{e})}\Big] \\
&= \mathbb{E}_{\mathbb{Q}}\Big[-\log \mathbb{P}_\theta(\mathbf{u}_T) + \sum_{t=1}^{T}\log \frac{\mathbb{Q}(\mathbf{u}_t|\mathbf{u}_{t-1})}{\mathbb{P}_\theta(\mathbf{u}_{t-1}|\mathbf{u}_t,\mathbf{e})}\Big] \\
&= \mathbb{E}_{\mathbb{Q}}\Big[-\log \mathbb{P}_\theta(\mathbf{u}_T) + \sum_{t=2}^{T}\log \frac{\mathbb{Q}(\mathbf{u}_t|\mathbf{u}_{t-1})}{\mathbb{P}_\theta(\mathbf{u}_{t-1}|\mathbf{u}_t,\mathbf{e})} + \log \frac{\mathbb{Q}(\mathbf{u}_1|\mathbf{u}_0)}{\mathbb{P}_\theta(\mathbf{u}_0|\mathbf{u}_1,\mathbf{e})}\Big] \\
&= \mathbb{E}_{\mathbb{Q}}\Big[-\log \mathbb{P}_\theta(\mathbf{u}_T) + \sum_{t=2}^{T}\log \Big(\frac{\mathbb{Q}(\mathbf{u}_{t-1}|\mathbf{u}_t,\mathbf{u}_0)}{\mathbb{P}_\theta(\mathbf{u}_{t-1}|\mathbf{u}_t,\mathbf{e})}\cdot\frac{\mathbb{Q}(\mathbf{u}_t|\mathbf{u}_0)}{\mathbb{Q}(\mathbf{u}_{t-1}|\mathbf{u}_0)}\Big) + \log \frac{\mathbb{Q}(\mathbf{u}_1|\mathbf{u}_0)}{\mathbb{P}_\theta(\mathbf{u}_0|\mathbf{u}_1,\mathbf{e})}\Big] \\
&= \mathbb{E}_{\mathbb{Q}}\Big[-\log \mathbb{P}_\theta(\mathbf{u}_T) + \sum_{t=2}^{T}\log \frac{\mathbb{Q}(\mathbf{u}_{t-1}|\mathbf{u}_t,\mathbf{u}_0)}{\mathbb{P}_\theta(\mathbf{u}_{t-1}|\mathbf{u}_t,\mathbf{e})} + \sum_{t=2}^{T}\log \frac{\mathbb{Q}(\mathbf{u}_t|\mathbf{u}_0)}{\mathbb{Q}(\mathbf{u}_{t-1}|\mathbf{u}_0)} + \log \frac{\mathbb{Q}(\mathbf{u}_1|\mathbf{u}_0)}{\mathbb{P}_\theta(\mathbf{u}_0|\mathbf{u}_1,\mathbf{e})}\Big] \\
&= \mathbb{E}_{\mathbb{Q}}\Big[-\log \mathbb{P}_\theta(\mathbf{u}_T) + \sum_{t=2}^{T}\log \frac{\mathbb{Q}(\mathbf{u}_{t-1}|\mathbf{u}_t,\mathbf{u}_0)}{\mathbb{P}_\theta(\mathbf{u}_{t-1}|\mathbf{u}_t,\mathbf{e})} + \log \frac{\mathbb{Q}(\mathbf{u}_T|\mathbf{u}_0)}{\mathbb{Q}(\mathbf{u}_1|\mathbf{u}_0)} + \log \frac{\mathbb{Q}(\mathbf{u}_1|\mathbf{u}_0)}{\mathbb{P}_\theta(\mathbf{u}_0|\mathbf{u}_1,\mathbf{e})}\Big] \\
&= \mathbb{E}_{\mathbb{Q}}\Big[\log \frac{\mathbb{Q}(\mathbf{u}_T|\mathbf{u}_0)}{\mathbb{P}_\theta(\mathbf{u}_T)} + \sum_{t=2}^{T}\log \frac{\mathbb{Q}(\mathbf{u}_{t-1}|\mathbf{u}_t,\mathbf{u}_0)}{\mathbb{P}_\theta(\mathbf{u}_{t-1}|\mathbf{u}_t,\mathbf{e})} - \log \mathbb{P}_\theta(\mathbf{u}_0|\mathbf{u}_1,\mathbf{e})\Big] \\
&= \mathbb{E}_{\mathbb{Q}}[\underbrace{\mathrm{KL}(\mathbb{Q}(\mathbf{u}_T|\mathbf{u}_0)\parallel \mathbb{P}_\theta(\mathbf{u}_T))}_{L_T} + \sum_{t=2}^{T}\underbrace{\mathrm{KL}(\mathbb{Q}(\mathbf{u}_{t-1}|\mathbf{u}_t,\mathbf{u}_0)\parallel \mathbb{P}_\theta(\mathbf{u}_{t-1}|\mathbf{u}_t,\mathbf{e}))}_{L_{t-1}} \underbrace{-\log \mathbb{P}_\theta(\mathbf{u}_0|\mathbf{u}_1,\mathbf{e})}_{L_0}]
\end{aligned}
\tag{12}
$$

Combine Eq. 11 and Eq. 12, we obtain:

$$
\begin{aligned}
\mathcal{L}_{\text{CE}} = -\mathbb{E}_{\mathbb{Q}}\log \mathbb{P}_\theta(\mathbf{u}_0|\mathbf{e}) \leq \mathbb{E}_{\mathbb{Q}}\Bigg[&\underbrace{\mathrm{KL}(\mathbb{Q}(\mathbf{u}_T|\mathbf{u}_0)\parallel \mathbb{P}_\theta(\mathbf{u}_T))}_{\mathcal{L}_T}\underbrace{-\log \mathbb{P}_\theta(\mathbf{u}_0|\mathbf{u}_1,\mathbf{e})}_{\mathcal{L}_0} \\
&+ \sum_{t=2}^{T}\underbrace{\mathrm{KL}(\mathbb{Q}(\mathbf{u}_{t-1}|\mathbf{u}_t,\mathbf{u}_0)\parallel \mathbb{P}_\theta(\mathbf{u}_{t-1}|\mathbf{u}_t,\mathbf{e})}_{\mathcal{L}_{t-1}}\Bigg].
\end{aligned}
\tag{13}
$$

□

To compute the KL divergence between probability measures $\mathrm{KL}(\mathbb{Q} \parallel \mathbb{P})$, we need to utilize a measure-theoretic definition of the KL divergence, which is stated in the following lemmas [2].

Lemma 1 (Measure Equivalence - The Feldman-Hájek Theorem). *Let $\mathbb{Q} = \mathcal{N}(\mathbf{m}_1, \mathbf{C}_1)$ and $\mathbb{P} = \mathcal{N}(\mathbf{m}_2, \mathbf{C}_2)$ be Gaussian measures on $\mathcal{H}$. They are equivalent if and only if $(i) : \mathbf{C}_1^{1/2}(\mathcal{H}) = \mathbf{C}_2^{1/2}(\mathcal{H}) = \mathcal{H}_0$, $(ii) : \mathbf{m}_1 - \mathbf{m}_2 \in \mathcal{H}_0$, and (iii) : The operator $(\mathbf{C}_1^{-1/2}\mathbf{C}_2^{1/2})(\mathbf{C}_1^{-1/2}\mathbf{C}_2^{1/2})^* - \mathbf{I}$ is a Hilbert-Schmidt operator on the closure $\overline{\mathcal{H}_0}$.*

Proof. Refer to the proof of Theorem 2.25 of Da Prato and Zabczyk [2]. □

Lemma 2 (The Radon-Nikodym Derivative). *Let* $\mathbb{Q} = \mathcal{N}(\mathbf{m}_1, \mathbf{C}_1)$ *and* $\mathbb{P} = \mathcal{N}(\mathbf{m}_2, \mathbf{C}_2)$ *be Gaussian measures on* $\mathcal{H}$*. If* $\mathbb{P}$ *and* $\mathbb{Q}$ *are equivalent and* $\mathbf{C}_1 = \mathbf{C}_2 = \mathbf{C}$*, then* $\mathbb{P}$*-a.s. the Radon-Nikodym derivative* $\mathrm{d}\mathbb{Q}/\mathrm{d}\mathbb{P}$ *is given by*

$$\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}(\mathbf{f}) = \exp\left[\left\langle \mathbf{C}^{-1/2}\left(\mathbf{m}_1 - \mathbf{m}_2\right), \mathbf{C}^{-1/2}\left(\mathbf{f} - \mathbf{m}_2\right)\right\rangle - \tfrac{1}{2}\|\mathbf{C}^{-1/2}(\mathbf{m}_1 - \mathbf{m}_2)\|^2\right] \forall \mathbf{f} \in \mathcal{H}. \tag{14}$$

Proof. Refer to the proof of Theorem 2.23 of Da Prato and Zabczyk [2]. □

Lemma 1 states the three conditions under which two Gaussian measures are equivalent. Lemma 2 is the consequence of the Feldman-Hájek theorem, providing the Radon-Nikodym derivative formula when dealing with Gaussian measures on $\mathcal{H}$.

To train the diffusion model in functional space we have to minimize the upper bound of Proposition 1, which requires us to compute the KL divergence between the measures $\mathbb{Q}, \mathbb{P}$. In order to satisfy Lemma 1, which will enable us to use Lemma 2 to compute the KL divergence, we make the following assumption:

Assumption 1 *Let* $\mathbb{Q} = \mathcal{N}(\tilde{\mathbf{m}}_t(\mathbf{u}_t, \mathbf{u}_0), \tilde{\beta}_t\mathbf{C})$ *and* $\mathbb{P}_\theta = \mathcal{N}(\mathbf{m}_\theta(\mathbf{u}_t, \mathbf{e}, t), \tilde{\beta}_t\mathbf{C})$ *be Gaussian measures on* $\mathcal{H}$*. With a conditional component* $\mathbf{e}$*, which can be an element of finite-dimensional space* $\mathbb{R}^d$ *or Hilbert space* $\mathcal{H}$*, there exists a parameter set* θ *such that the difference in mean elements of the two measures falls within the scaled covariance space:*

$$\tilde{\mathbf{m}}_t(\mathbf{u}_t, \mathbf{u}_0) - \mathbf{m}_\theta(\mathbf{u}_t, \mathbf{e}, t) \in (\tilde{\beta}_t\mathbf{C})^{1/2}(\mathcal{H}). \tag{15}$$

By making this assumption we satisfy all three conditions of Lemma 1: (i) : $\mathbf{C}_1^{1/2}(\mathcal{H}) = \mathbf{C}_2^{1/2}(\mathcal{H}) = (\tilde{\beta}_t\mathbf{C})^{1/2}(\mathcal{H}) = \mathcal{H}_0$; (ii) : $\mathbf{m}_1 - \mathbf{m}_2 \in \mathcal{H}_0$ is directly satisfied from Assumption 1; (iii) : $(\mathbf{C}_1^{-1/2}\mathbf{C}_2^{1/2})(\mathbf{C}_1^{-1/2}\mathbf{C}_2^{1/2})^* - \mathbf{I} = \mathbf{I} - \mathbf{I}$ is the zero operator, which is trivially a Hilbert-Schmidt operator as its Hilbert-Schmidt norm is 0. As a consequence, $\mathbb{Q}$ and $\mathbb{P}$ are equivalent, allowing us to utilize the Radon-Nikodym derivative from Lemma 2.

Theorem 1 (Conditional Diffusion Optimality in Function Space). *Given the specified conditions in Assumption 1, the minimization of the learning objective in Proposition 1 is equivalent to obtaining the parameter set* θ^* *that is the solution to the problem*

$$\theta^* = \arg\min_\theta \mathbb{E}_{\mathbf{u}_0 \sim \mathbb{Q}_{\text{data}}, t \sim [1,T]} \lambda_t \left|\left| \mathbf{C}^{-1/2}\left(\mathbf{A}\boldsymbol{\xi} - \boldsymbol{\xi}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{A}\mathbf{u}_0 + \sqrt{1-\bar{\alpha}_t}\mathbf{A}\boldsymbol{\xi}, \mathbf{e}, t)\right)\right|\right|_{\mathcal{H}}^2, \tag{16}$$

where $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$*,* $\mathbf{A} : \mathcal{H} \to \mathcal{H}$ *denotes a smoothing operator,* $\mathbf{e} \in (\mathbb{R}^d \cup \mathcal{H})$ *is a conditional component,* $\boldsymbol{\xi}_\theta : \{1, 2, \ldots, T\} \times (\mathbb{R}^d \cup \mathcal{H}) \times \mathcal{H} \to \mathcal{H}$ *is a parameterized mapping,* $\lambda_t = \beta_t^2/2\tilde{\beta}_t(1-\beta_t)(1-\bar{\alpha}_t) \in \mathbb{R}$ *is a time-dependent constant.*

Proof. Under Assumption 1, we are now able to use the Radon-Nikodym derivative to compute the KL divergence:

$$\begin{aligned}\mathrm{KL}\left[\mathbb{Q} \parallel \mathbb{P}\right] &= \int_{\mathcal{H}} \log \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}(\mathbf{f})\ \mathrm{d}\mathbb{Q}(\mathbf{f}) \\ &= -\frac{1}{2}\|\mathbf{C}^{-1/2}(\mathbf{m}_1 - \mathbf{m}_2)\|_{\mathcal{H}}^2 + \int_{\mathcal{H}} \left\langle \mathbf{C}^{-1/2}\left(\mathbf{m}_1 - \mathbf{m}_2\right), \mathbf{C}^{-1/2}\left(\mathbf{f} - \mathbf{m}_2\right)\right\rangle\ \mathrm{d}\mathbb{Q}(\mathbf{f}).\end{aligned} \tag{17}$$

We now use spectral decomposition to compute the integral term. Let $\{\lambda_j, \mathbf{e}_j\}_{j=1}^{\infty}$ be the eigenvalues and eigenvectors of $\mathbf{C}$. The eigenvector of $\mathbf{C}$ form an orthonormal basis for $\mathcal{H}$ by the spectral theorem, as $\mathbf{C}$ is a self-adjoint compact operator. Hence, the second integral is:

$$\begin{aligned}&\int_{\mathcal{H}} \left\langle \mathbf{C}^{-1/2}\left(\mathbf{m}_1 - \mathbf{m}_2\right), \mathbf{C}^{-1/2}\left(\mathbf{f} - \mathbf{m}_2\right)\right\rangle\ \mathrm{d}\mathbb{Q}(f) \\ &= \int_{\mathcal{H}} \sum_{j=1}^{\infty} \left\langle \mathbf{m}_1 - \mathbf{m}_2, \mathbf{e}_j\right\rangle \left\langle \mathbf{f} - \mathbf{m}_2, \mathbf{e}_j\right\rangle \lambda_j^{-1}\ \mathrm{d}\mathbb{Q}(f) \\ &= \sum_{j=1}^{\infty} \lambda_j^{-1} \left\langle \mathbf{m}_1 - \mathbf{m}_2, \mathbf{e}_j\right\rangle \int_{\mathcal{H}} \left\langle \mathbf{f} - \mathbf{m}_2, \mathbf{e}_j\right\rangle\ \mathrm{d}\mathbb{Q}(f) \\ &= \sum_{j=1}^{\infty} \lambda_j^{-1} \left\langle \mathbf{m}_1 - \mathbf{m}_2, \mathbf{e}_j\right\rangle^2 \\ &= \left\langle \mathbf{C}^{-1/2}\left(\mathbf{m}_1 - \mathbf{m}_2\right), \mathbf{C}^{-1/2}\left(\mathbf{m}_1 - \mathbf{m}_2\right)\right\rangle.\end{aligned} \tag{18}$$

Combine Eq. 17 and Eq. 18, we obtain:

$$\mathrm{KL}\left[\mathbb{Q} \parallel \mathbb{P}\right] = \frac{1}{2}\|\mathbf{C}^{-1/2}(\mathbf{m}_1 - \mathbf{m}_2)\|_{\mathcal{H}}^2 \tag{19}$$

From Proposition 1, the KL divergence between Gaussian measures $\mathbb{Q}$ and $\mathbb{P}$ now becomes:

$$\begin{aligned}L_{t-1} &= \mathrm{KL}\left[\mathbb{Q}(\mathbf{u}_{t-1}|\mathbf{u}_t, \mathbf{u}_0) \parallel \mathbb{P}_\theta(\mathbf{u}_{t-1}|\mathbf{u}_t, \mathbf{e})\right] \\ &= \frac{1}{2}\|(\tilde{\beta}_t\mathbf{C})^{-1/2}(\tilde{\mathbf{m}}_t(\mathbf{u}_t, \mathbf{u}_0) - \mathbf{m}_\theta(\mathbf{u}_t, \mathbf{e}, t))\|_{\mathcal{H}}^2\end{aligned} \tag{20}$$

Our model must predict the mean function $\tilde{\mathbf{m}}_t(\mathbf{u}_t, \mathbf{u}_0)$. Recall that we got the expression of $\tilde{\mathbf{m}}_t(\mathbf{u}_t, \mathbf{u}_0)$ and $\mathbf{u}_0$ depending on $\mathbf{u}_t$:

$$\tilde{\mathbf{m}}_t(\mathbf{u}_t, \mathbf{u}_0) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1 - \bar{\alpha}_t}\mathbf{A}\mathbf{u}_0 + \frac{\sqrt{1 - \beta_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t}\mathbf{u}_t. \tag{21}$$

$$\mathbf{A}\mathbf{u}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{u}_t - \sqrt{1 - \bar{\alpha}_t}\mathbf{A}\boldsymbol{\xi}\right) \quad ; \text{where } \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}) \tag{22}$$

Combine these two expressions, we have:

$$\tilde{\mathbf{m}}_t(\mathbf{u}_t, \mathbf{u}_0) = \frac{1}{\sqrt{1 - \beta_t}}\left(\mathbf{u}_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}}\mathbf{A}\boldsymbol{\xi}\right) \tag{23}$$

Thus, we parameterize the variational mean via:

$$\mathbf{m}_\theta(\mathbf{u}_t, \mathbf{e}, t) = \frac{1}{\sqrt{1-\beta_t}}\left(\mathbf{u}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\xi}_\theta(\mathbf{u}_t, \mathbf{e}, t)\right) \tag{24}$$

Finally, plugging Eq. 23 and Eq. 24 into L_{t-1}, we obtain:

$$\begin{aligned} L_{t-1} &= \frac{1}{2}\left\|(\tilde{\beta}_t\mathbf{C})^{-1/2}\left(\frac{1}{\sqrt{1-\beta_t}}\frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\mathbf{A}\boldsymbol{\xi} - \frac{1}{\sqrt{1-\beta_t}}\frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\xi}_\theta(\mathbf{u}_t, \mathbf{e}, t)\right)\right\|_{\mathcal{H}}^2 \\ &= \frac{\beta_t^2}{2\tilde{\beta}_t(1-\beta_t)(1-\bar{\alpha}_t)}\left\|\mathbf{C}^{-1/2}\left(\mathbf{A}\boldsymbol{\xi} - \boldsymbol{\xi}_\theta(\mathbf{u}_t, \mathbf{e}, t)\right)\right\|_{\mathcal{H}}^2 \\ &= \frac{\beta_t^2}{2\tilde{\beta}_t(1-\beta_t)(1-\bar{\alpha}_t)}\left\|\mathbf{C}^{-1/2}\left(\mathbf{A}\boldsymbol{\xi} - \boldsymbol{\xi}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{A}\mathbf{u}_0 + \sqrt{1-\bar{\alpha}_t}\mathbf{A}\boldsymbol{\xi}, \mathbf{e}, t)\right)\right\|_{\mathcal{H}}^2 \end{aligned} \tag{25}$$

□

B Experiments

B.1 Long-range dependencies

We obtained the patch-based large-image model of Graikos et al. [3] directly from the authors and tried to apply it to synthesize images larger than 1024×1024 pixels. The overreliance of the patch-based model on the local descriptors (patch SSL embeddings) leads to the loss of large-scale structures and fails to capture long-range dependencies across the image. As a qualitative example (Figure 1), we get a reference image of size 2048×2048 pixels from TCGA-BRCA and extract embeddings in an attempt to generate a variation of it using our model and the patch-based model of [3]. As illustrated, ∞-Brush retains large-scale structures (such as clearly-separated clusters of cells) that can span multiple patches, in comparison to the image generated from [3].

B.2 Zero-shot classification

Following the experiment of [3], we generate images from a pre-defined set of four classes: benign tissue, in-situ, invasive carcinoma, and normal tissue. We use a VLM (Quilt) as a zero-shot classifier and compute the confusion matrix (CM). Figure 2 shows that ∞-Brush generates images semantically aligned with the text prompts.

B.3 Application of synthetic data on downstream task

As a practical application, we double the number of training images of the BACH dataset by synthesizing images using real data embedding and evaluating the test set. Table 1 shows a significant accuracy boost from these synthetic images.

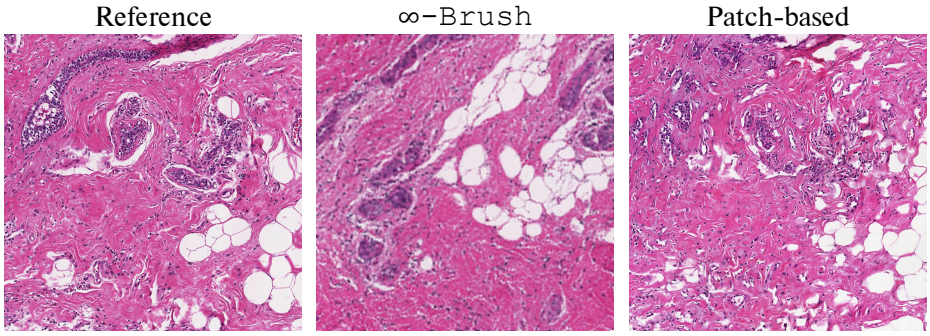

Fig. 1: Long-range dependencies comparison between our ∞-**Brush** and patched-based method [3]. ∞-**Brush** retains large-scale structures (such as clearly-separated clusters of cells) that can span multiple patches in comparison to the image generated from [3].

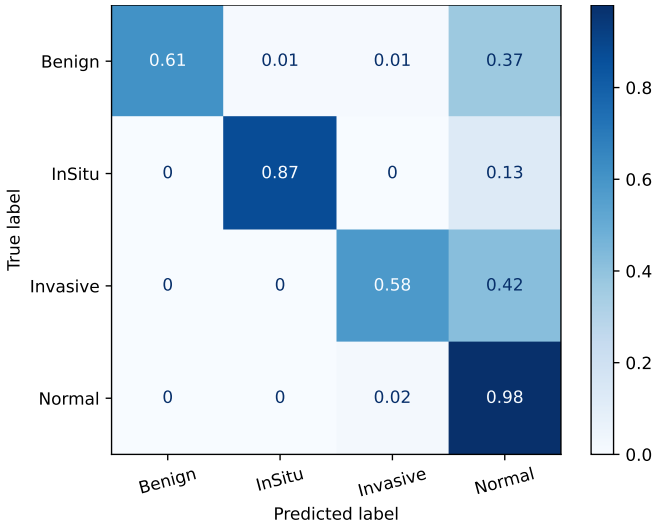

Fig. 2: Confusion matrix of zero-shot classification of generated images.

Table 1: Synthetic data improves the accuracy significantly on the BACH test set.

Training Data	Test Acc
Real	79 %
Real + synthetic	**83 %**

B.4 Ablation study on % of pixels for training

We compare our model when training on 256*256 (0.4%) *vs.* 512*512 pixels (1.6%). Figure 3 shows that training with more pixels improves performance. Our model efficiently uses 0.4% of pixels compared to 25% of ∞-Diff's due to

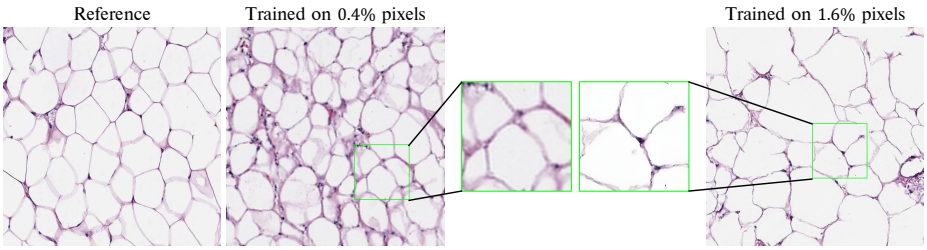

Fig. 3: Ablation on % pixels for training and zoomed-in views.

the incorporation of coordinate embedding in CANO, functioning as positional embedding.

B.5 Qualitative results

In Figure 4 and Figure 5, we illustrate the generated very large (4096×4096) and large (1024×1024) images of TCGA-BRCA [1] dataset. We also show synthesized satellite images at 2048×2048 and 1024×1024 resolutions in Figure 6. Qualitative results show that given a single embedding vector of a downsampled 256×256 real image, ∞-Brush can synthesize images of arbitrary resolutions up to 4096×4096 and preserve global structures of the reference image.

Figure 7 shows examples where the model did not successfully capture spatial structures and details from the reference images. This can be attributed to both the model and the conditioning used to represent the images.

References

1. Cancer Genome Atlas Research Network, J., et al.: The cancer genome atlas pan-cancer analysis project. Nat. Genet **45**(10), 1113–1120 (2013) 1, 9
2. Da Prato, G., Zabczyk, J.: Stochastic Equations in Infinite Dimensions. Encyclopedia of Mathematics and its Applications, Cambridge University Press, 2 edn. (2014) 4, 5
3. Graikos, A., Yellapragada, S., Le, M.Q., Kapse, S., Prasanna, P., Saltz, J., Samaras, D.: Learned representation-guided diffusion models for large-image generation. arXiv preprint arXiv:2312.07330 (2023) 1, 7, 8
4. Hoogeboom, E., Salimans, T.: Blurring diffusion models. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=OjDkC57x5sz 2
5. Rissanen, S., Heinonen, M., Solin, A.: Generative modelling with inverse heat dissipation. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=4PJUBT9f2Ol 2
6. USGS: National agriculture imagery program (NAIP) (2023), https://www.usgs.gov/centers/eros/science/usgs-eros-archive-aerial-photography-national-agriculture-imagery-program-naip 1

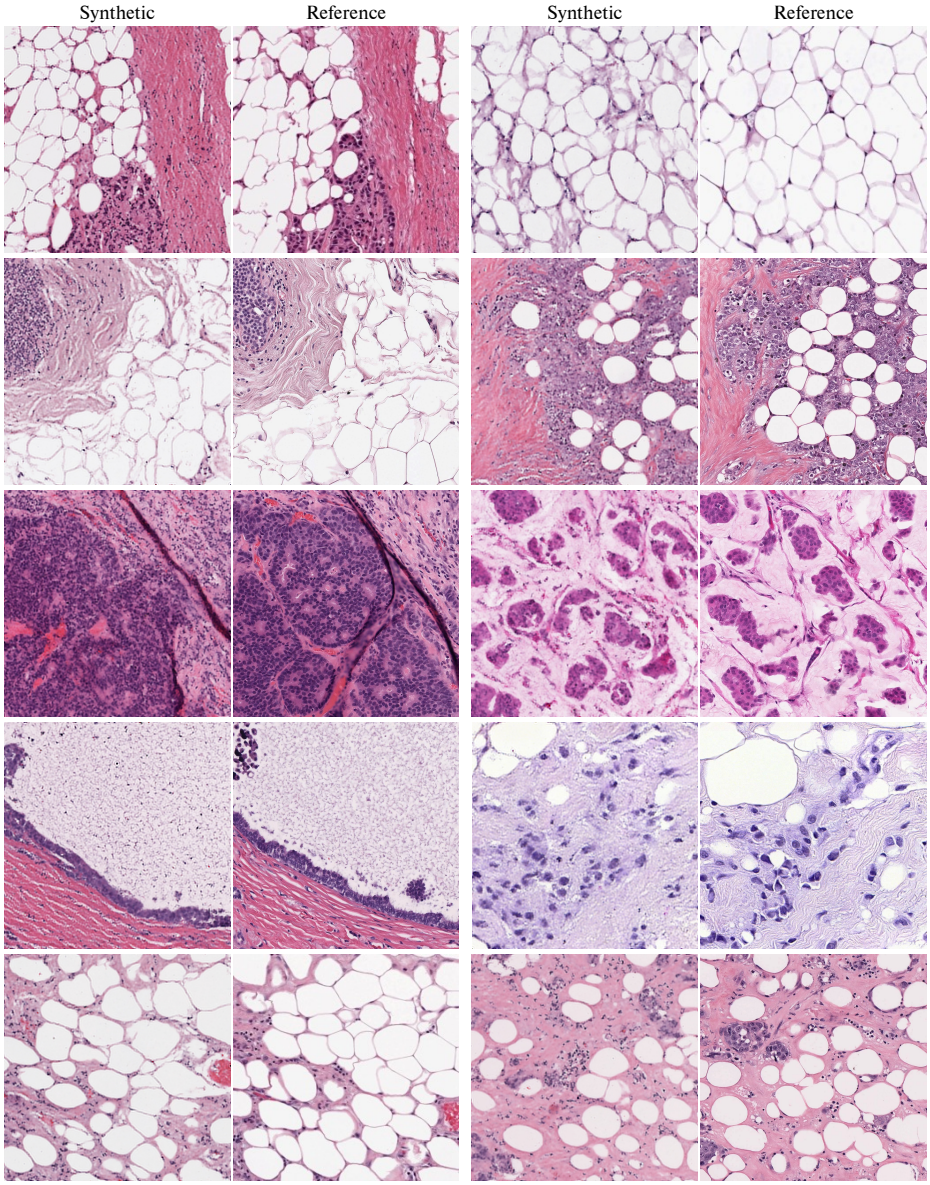

Fig. 4: Very large (4096×4096) images generated from ∞-`Brush`, and the corresponding reference real images used to generate them. Given a single embedding vector of a downsampled 256×256 real image, ∞-`Brush` can synthesize images of up to 4096×4096 and preserve global structures of the reference image.

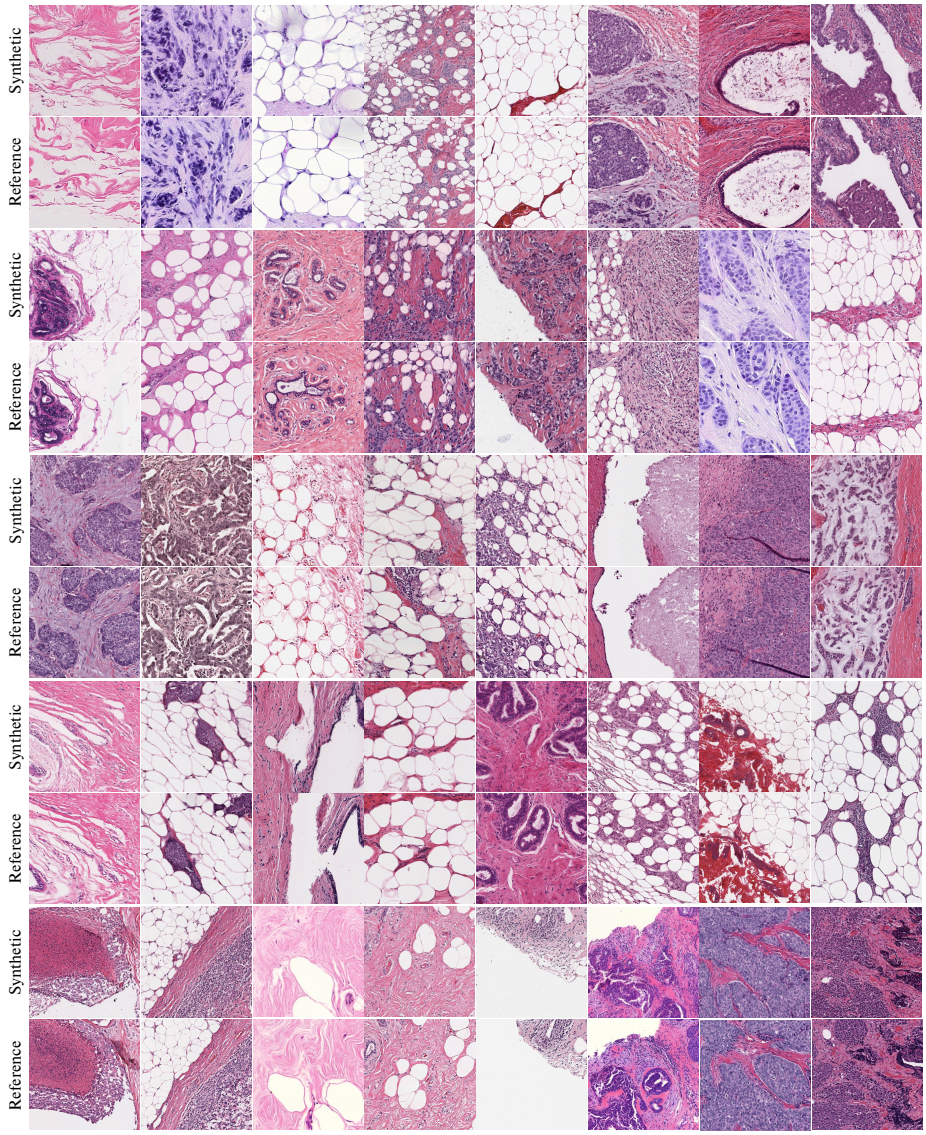

Fig. 5: Large (1024×1024) images generated from ∞-Brush, and the corresponding reference real images used to generate them. Given a single embedding vector of a downsampled 256×256 real image, ∞-Brush can synthesize images at arbitrary resolutions and preserve global structures of the reference image.

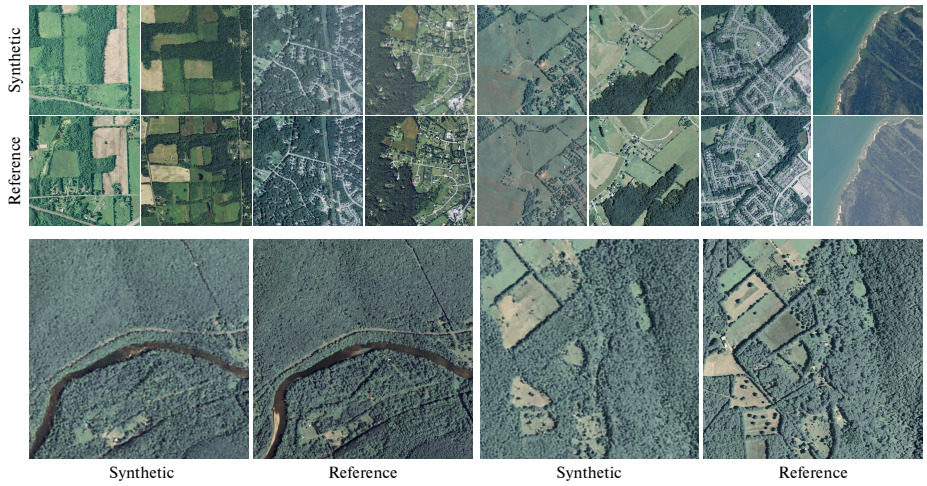

Fig. 6: Satellite large (1024×1024 and 2048×2048) images generated from ∞-`Brush`, and the corresponding reference real images used to generate them. Given a single embedding vector of a downsampled 256×256 real image, ∞-`Brush` can synthesize images at arbitrary resolutions and preserve global structures of the reference image.

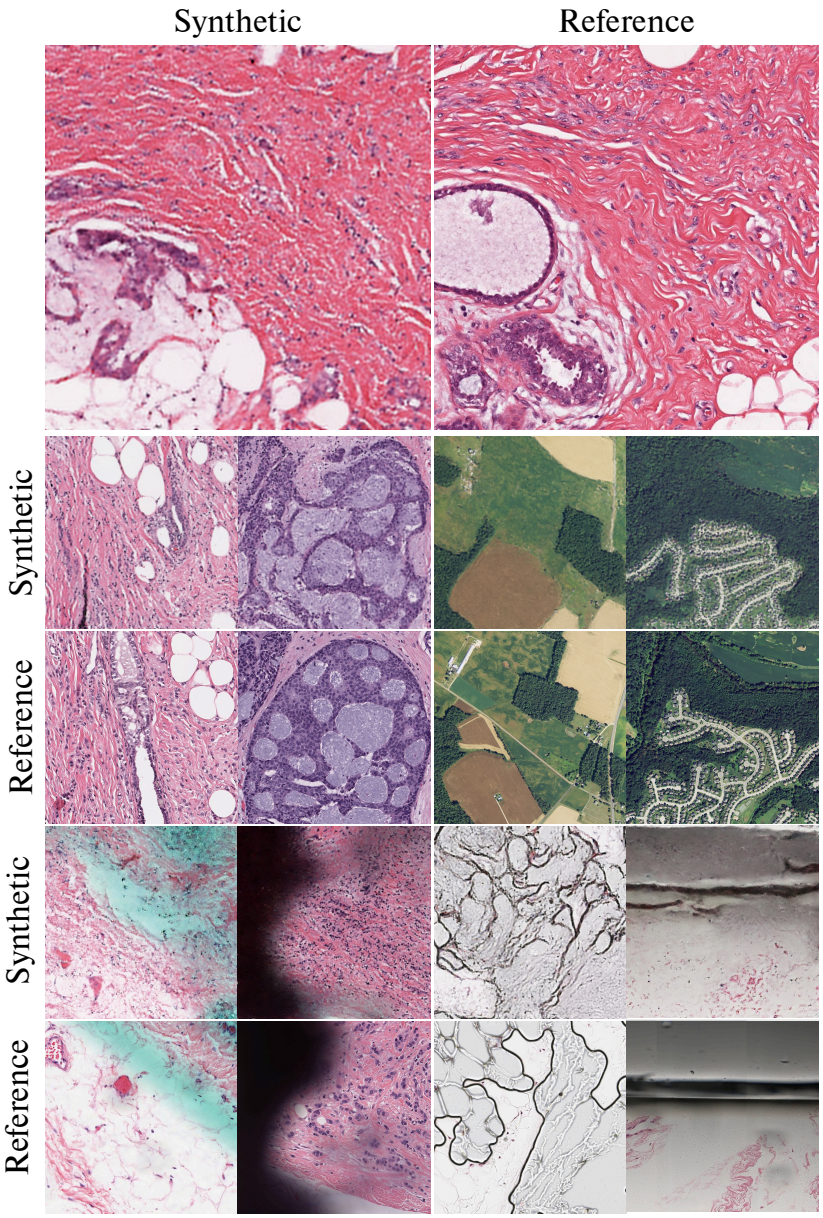

Fig. 7: Uncurated (4096 × 4096 and 2048 × 2048) images generated from ∞-Brush, and the corresponding reference real images used to generate them. Our model fails to capture spatial structure and details in specific regions of reference images (top 3 rows). In the last 2 rows, it shows that our model fails to controllably synthesize images due to bad conditioning information.