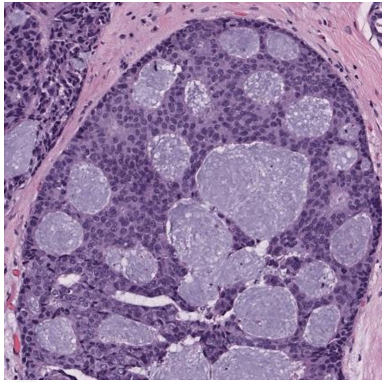
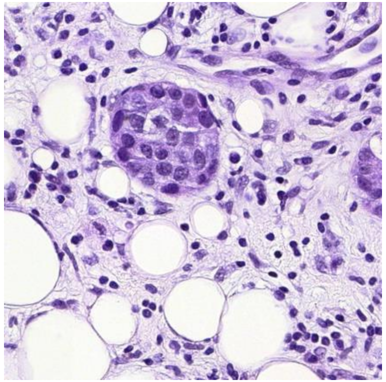
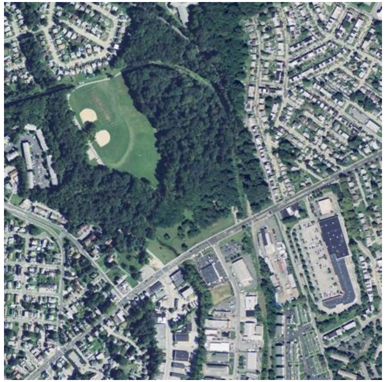
∞-Brush 🎨: Controllable Large Image Synthesis with Diffusion Models in Infinite Dimensions

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

ECCV 2024

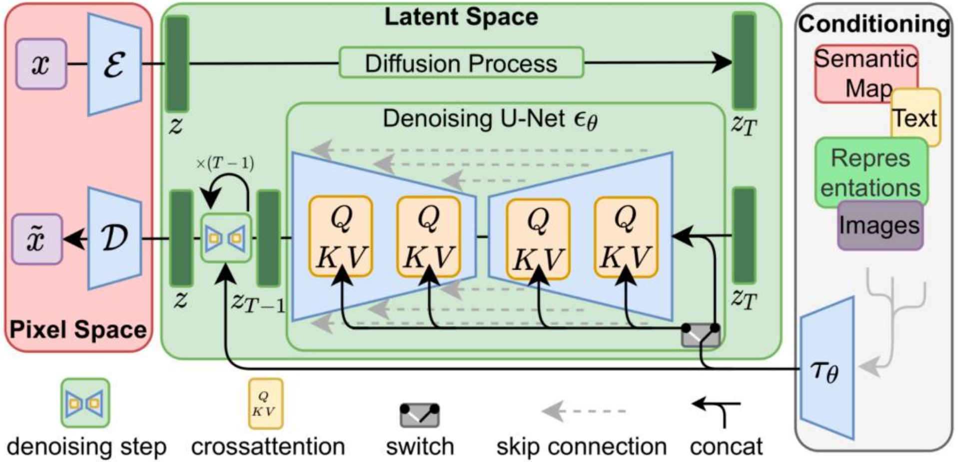
Introduction

- Diffusion models can synthesize diverse and complex data, e.g. images and videos.
- Still difficult to generate high-resolution images, especially when conditioning on intricate, domain-specific information, e.g. histopathology and satellite images.


Current Approaches – Finite-dimensional Diffusion

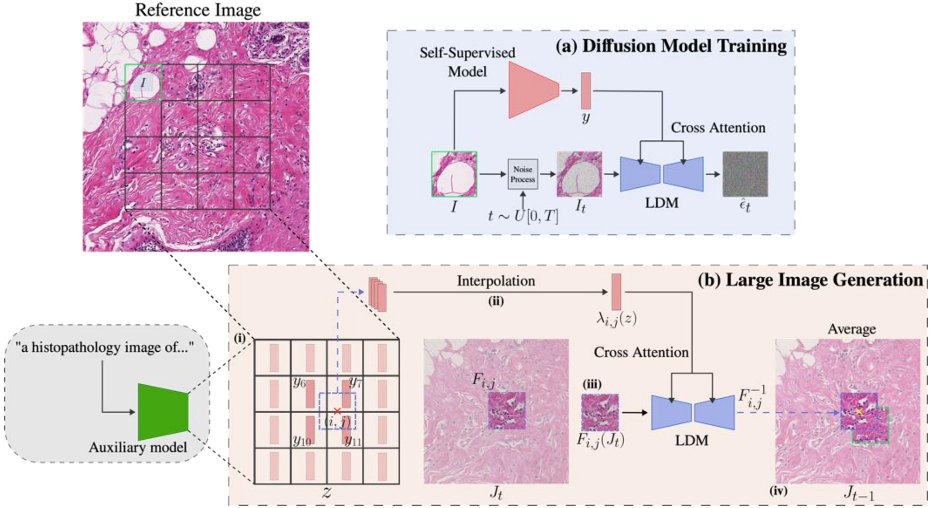
- Conditional diffusion models in finite dimensions: *e.g.* Stable Diffusion-XL, Matryoshka Diffusion, …, can generate images at fixed resolution (1024×1024).
- As resolution increases, computational resources scale quadratically.



Rombach et al., "High-Resolution Image Synthesis With Latent Diffusion Models", CVPR 2022

Current Approaches – Patch-based Diffusion

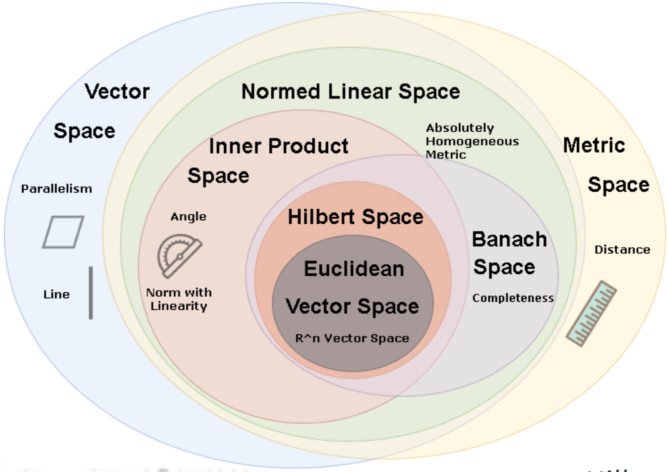
- Splits large image generation into smaller segments and perform large image synthesis via outpainting algorithm.
- While more computationally efficient and produces realistic larger images, it falls short of capturing long-range dependency.

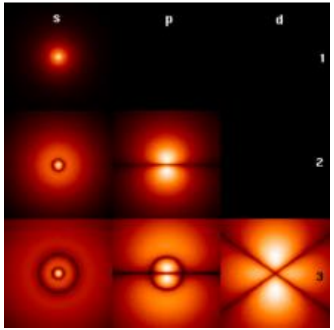


A. Graikos, S. Yellapragada, M.Q. Le, S. Kapse, P. Prasanna, J. Saltz, D. Samaras, "Learned representation-guided diffusion models for large-image generation", CVPR 2024

Current Approaches – Infinite-dimensional Diffusion

- Represents images as funtions in Hilbert space $\mathcal{H}$, can synthesize images at arbitrary resolution while training on fixed-size inputs.
- Current infinite-dimensional diffusion models cannot be conditioned for controllable image generation.

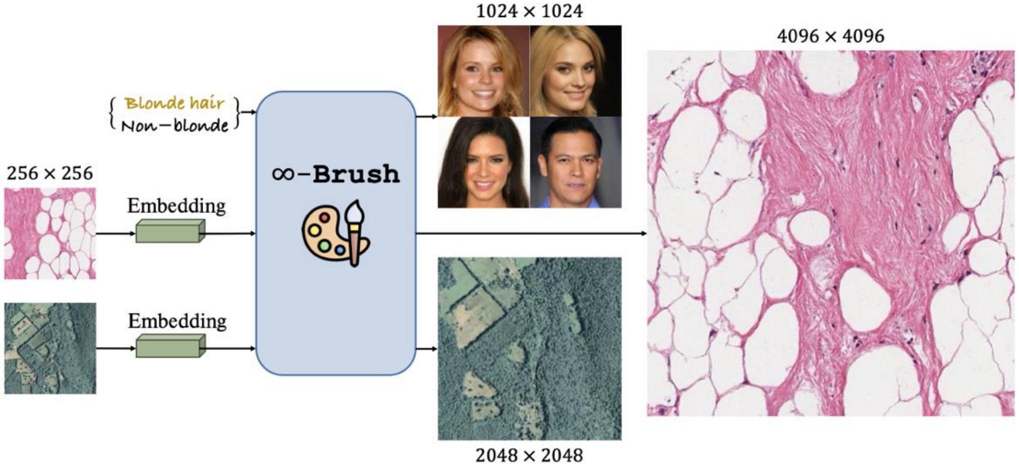




Hilbert space and examples, Wikipedia

Our Proposal: ∞-Brush

- Propose a cross-attention neural operator in function space, to incorporate external information during image generation.
- Build a conditional denoiser in function space as part of ∞-Brush, the first conditional diffusion model in function space.
- The first method to controllably synthesize images at arbitrary resolutions up to 4096×4096 pixels.



Preliminaries – Notation and Data

- A dataset of the form $\mathcal{D} = \{(\mathbf{u}_k, \mathbf{e}_k)\}_{1\leq k\leq D}$, where each $\mathbf{u}_j \in \mathcal{H}$ is an i.i.d. draw from an unknown probability measure $\mathbb{Q}_{\mathrm{data}}$ on $\mathcal{H}$, and $\mathbf{e}_j$ is a control component of function $\mathbf{u}_j$.
- It is difficult to represent the function directly, we discretize it on the mesh $\mathbf{x}_j = \{\mathbf{x}_j^{(i)}\}_{1\leq i\leq N} \subset \mathcal{X}$, with discretized observations $\{\mathbf{u}_j(\mathbf{x}_j^{(i)})\}_{1\leq i\leq N}$, being the output of function $\mathbf{u}_j$ at the $i\text{-th}$ observation point.

Preliminaries – Gaussian Measures on Hilbert Spaces

Let $\mathbb{Q}$ be a probability measure on $(\mathcal{H}, \mathcal{B}(\mathcal{H}))$. If $\mathbb{Q}$ is Gaussian, then there exists a mean element $\mathbf{m} \in \mathcal{H}$ and a covariance operator $\mathbf{C} : \mathcal{H} \to \mathcal{H}$, such that

$$\int_{\mathcal{H}} \langle \mathbf{u}, \mathbf{x} \rangle \mathbb{Q}(\mathrm{d}\mathbf{x}) = \langle \mathbf{m}, \mathbf{u} \rangle, \quad \forall \mathbf{u} \in \mathcal{H}, \tag{1}$$

$$\int_{\mathcal{H}} \langle \mathbf{u}_1, \mathbf{x} - \mathbf{m} \rangle \langle \mathbf{u}_2, \mathbf{x} - \mathbf{m} \rangle \mathbb{Q}(\mathrm{d}\mathbf{x}) = \langle \mathbf{C}\mathbf{u}_1, \mathbf{u}_2 \rangle, \quad \forall \mathbf{u}_1, \mathbf{u}_2 \in \mathcal{H}. \tag{2}$$

Preliminaries – Neural Operators

- A type of neural network tailored to learn mappings between infinite-dimensional function spaces.
- In diffusion models in infite dimensions, a denoiser is parameterized by a neural operator $\mathcal{G}_\theta : \mathcal{U}^* \to \mathcal{U}$, learns to map from noisy function space to denoised function space.
- Include multiple operator layers $\mathbf{v}_0 \mapsto \mathbf{v}_1 \mapsto \cdots \mapsto \mathbf{v}_L$, where layer $\mathbf{v}_l \mapsto \mathbf{v}_{l+1}$ is built on a local linear operator, a non-local integral kernel operator and a bias function:

$$\mathbf{v}_{l+1}(\mathbf{x}^{(i)}) = \sigma_{l+1}\left(W_l \mathbf{v}_l(\mathbf{x}^{(i)}) + (\mathcal{K}_l(\mathbf{u};\phi)\mathbf{v}_l)(\mathbf{x}^{(i)}) + b_l(\mathbf{x}^{(i)})\right)$$

∞-Brush 🎨 - Conditional Diffusion Models in Function Space

- **Forward process:** gradually perturbs the probability measure $\mathbb{Q}_0 = \mathbb{Q}_{\text{data}}$ towards a Gaussian measure $\mathcal{N}(\mathbf{m}, \mathbf{C})$

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

A smoothing operator $\mathbf{A} : \mathcal{H} \to \mathcal{H}$, *e.g.* a truncated Gaussian kernel, is applied to get a smoother function representation.

∞-Brush 🎨 - Conditional Diffusion Models in Function Space

- **Reverse process:** approximate posterior measures with a variational family of measures on $\mathcal{H}$ and use conditional embedding $\mathbf{e}$ to control the generation process

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

∞-Brush 🎨 - Conditional Diffusion Models in Function Space

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

$$\mathcal{L}_{\mathrm{CE}} = -\mathbb{E}_{\mathbb{Q}} \log \mathbb{P}_\theta(\mathbf{u}_0|\mathbf{e}) \leq \mathbb{E}_{\mathbb{Q}} \Bigg[\underbrace{\mathrm{KL}(\mathbb{Q}(\mathbf{u}_T|\mathbf{u}_0) \parallel \mathbb{P}_\theta(\mathbf{u}_T))}_{\mathcal{L}_T} \underbrace{- \log \mathbb{P}_\theta(\mathbf{u}_0|\mathbf{u}_1, \mathbf{e})}_{\mathcal{L}_0} + \sum_{t=2}^{T} \underbrace{\mathrm{KL}(\mathbb{Q}(\mathbf{u}_{t-1}|\mathbf{u}_t, \mathbf{u}_0) \parallel \mathbb{P}_\theta(\mathbf{u}_{t-1}|\mathbf{u}_t, \mathbf{e}))}_{\mathcal{L}_{t-1}} \Bigg]. \tag{11}$$

Proof. Please refer to the Supplementary Material for the full proof. □

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* $\overline{\mathcal{H}_0}$.

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*

$$\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}(\mathbf{f}) = \exp\left[\left\langle \mathbf{C}^{-1/2}\left(\mathbf{m}_1 - \mathbf{m}_2\right), \mathbf{C}^{-1/2}\left(\mathbf{f} - \mathbf{m}_2\right)\right\rangle - \frac{1}{2}\|\mathbf{C}^{-1/2}(\mathbf{m}_1 - \mathbf{m}_2)\|^2\right] \forall \mathbf{f} \in \mathcal{H}. \tag{12}$$

Proof. The proof of both lemmas is in the Supplementary Material. □

∞-Brush 🎨- Conditional Diffusion Models in Function Space

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:*

$$\tilde{\mathbf{m}}_t(\mathbf{u}_t, \mathbf{u}_0) - \mathbf{m}_\theta(\mathbf{u}_t, \mathbf{e}, t) \in (\tilde{\beta}_t \mathbf{C})^{1/2}(\mathcal{H}). \tag{13}$$

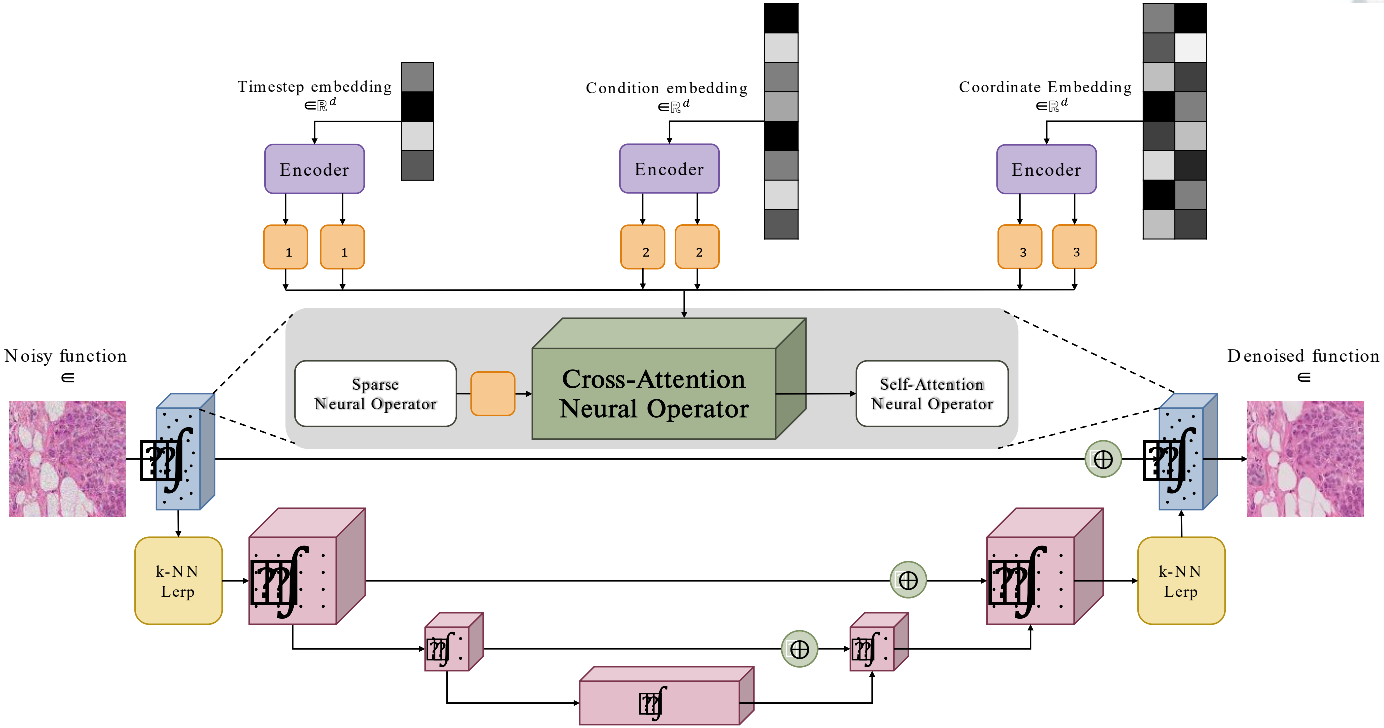
Theorem 1 (Conditional Diffusion Optimality in Function Space). *Given the specified conditions in Assumption 1, the minimization of the learning objective in Proposition 1 is equivalent to obtaining the parameter set θ^* that is the solution to the problem*

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{\mathbf{u}_0 \sim \mathbb{Q}_{\mathrm{data}}} \lambda_t \left|\left| \mathbf{C}^{-1/2} \left(\mathbf{A}\boldsymbol{\xi} - \boldsymbol{\xi}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{A}\mathbf{u}_0 + \sqrt{1-\bar{\alpha}_t}\mathbf{A}\boldsymbol{\xi}, \mathbf{e}, t) \right) \right|\right|_{\mathcal{H}}^2, \tag{14}$$

where $\boldsymbol{\xi} \sim \mathcal{N}(0, \mathbf{C})$, $\mathbf{A} : \mathcal{H} \to \mathcal{H}$ 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. Please refer to the Supplementary Material for the full proof. □

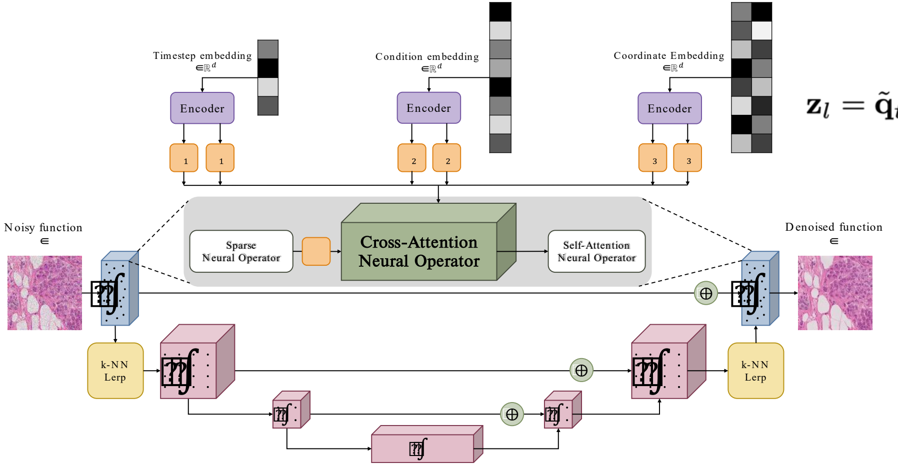
Conditional Denoiser with Cross-Attention Neural Operators



The sparse level utilizes a sparse neural operator, a cross-attention neural operator, and a self-attention neural operator, focusing on capturing fine-grained details. The grid level targets global information.

Conditional Denoiser with Cross-Attention Neural Operators

- The computational complexity of vanilla attention is quadratic $\mathcal{O}(N^2 d)$.
- We propose a cross-attention neural operator of linear complexity w.r.t. N
- Suppose we have L conditional embeddings $\{Y_l \in \mathbb{R}^{N_l \times d}\}_{1 \leq l \leq L}$, we first compute queries $Q = (\mathbf{q}_i)$, keys $K_l = (\mathbf{k}_i^l) = Y_l W_k$, and values $V_l = (\mathbf{v}_i^l) = Y_l W_v$



$$\mathbf{z}_l = \tilde{\mathbf{q}}_t + \frac{1}{L}\sum_{l=1}^{L}\sum_{i=1}^{N_l}\alpha_t^l\left(\tilde{\mathbf{q}}_t\cdot\tilde{\mathbf{k}}_i^l\right)\mathbf{v}_i^l = \tilde{\mathbf{q}}_t + \frac{1}{L}\sum_{l=1}^{L}\alpha_t^l\tilde{\mathbf{q}}_t\cdot\left(\sum_{i=1}^{N_l}\tilde{\mathbf{k}}_i^l\odot\mathbf{v}_i^l\right)$$

- The complexity is $\mathcal{O}((N+\sum_l N_l)d^2)$, which is linear w.r.t. N.

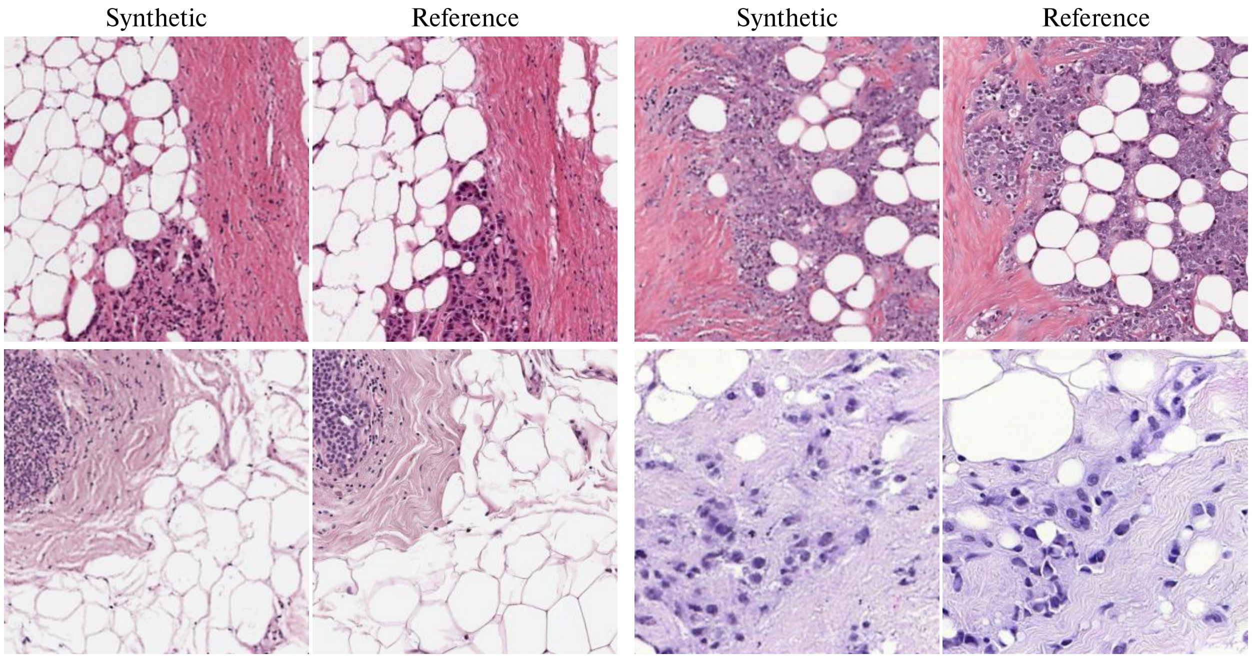
Experiments – Facial Attribute Conditional Generation

Large images (1024×1024) generated from our ∞-Brush, conditioned on the facial attribute blonde/non-blonde hair.

Table 1: The CLIP FID scores of our ∞-**Brush** model against ∞-Diff showcases our model's capability in conditionally generating celebrity faces on the CelebA-HQ dataset based on the facial attribute of hair color (blonde vs. non-blonde).

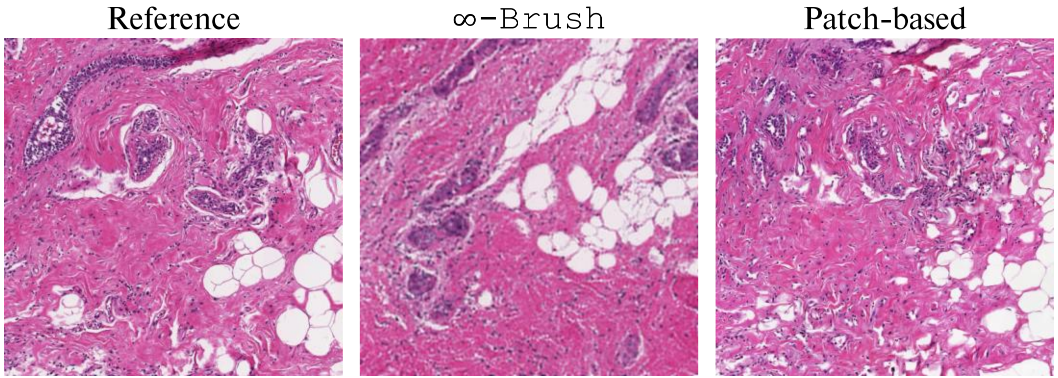
Dataset	# Images	Method	Training Config.	CLIP FID
CelebA-HQ (1024×1024)	30k	∞-Diff [2]	Unconditional	9.44
		∞-**Brush**	blonde vs. non-blonde hair	**8.38**

Experiments – Controllable Very Large Image Generation



Very large images (4096×4096) generated from ∞-Brush 🎨, and the corresponding reference real images used to generate them.

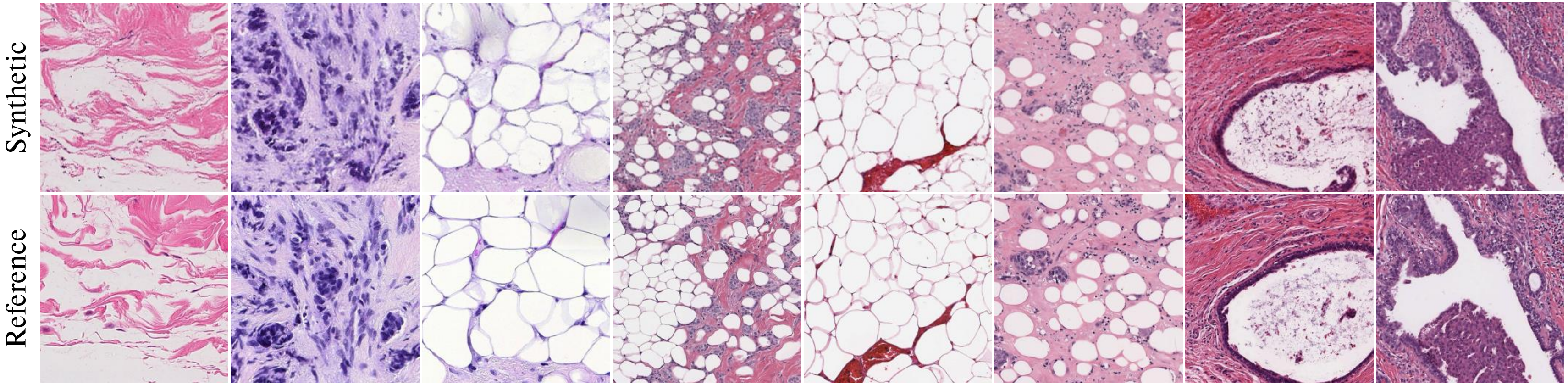
Experiments – Controllable Very Large Image Generation



∞-Brush 🎨 retains large-scale structures that can span multiple patches compared to the image generated from patch-based method.

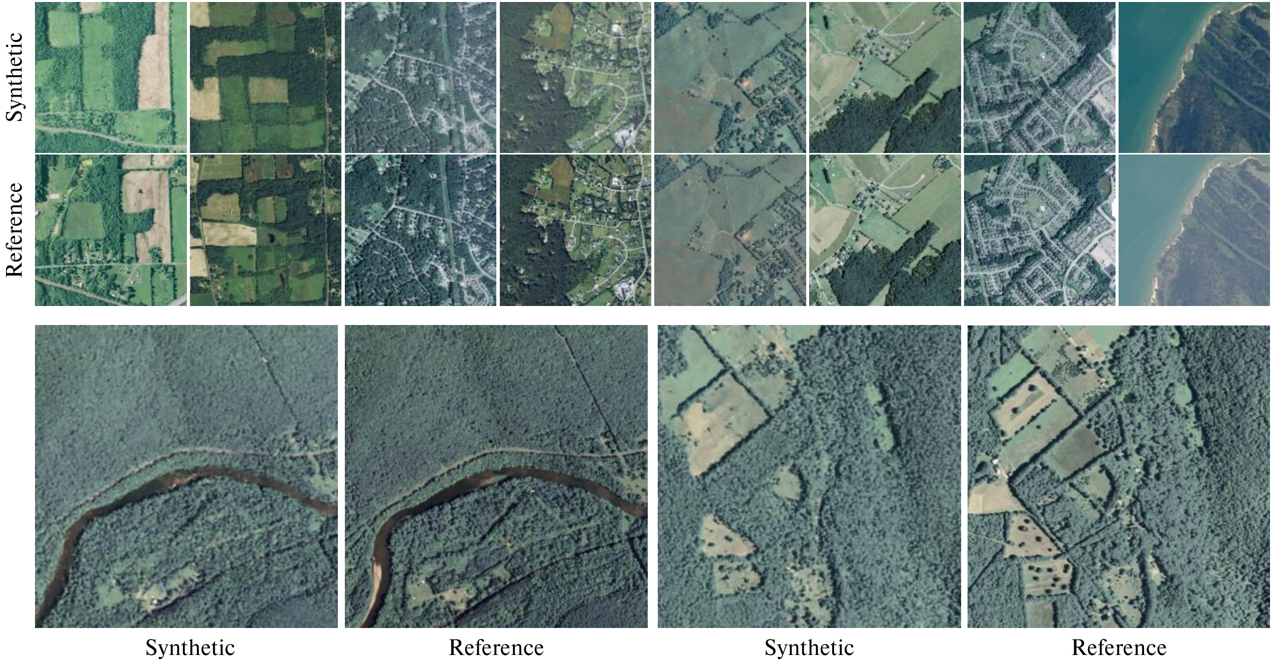
Dataset	# Images	Method	Training Config.	CLIP FID	Crop FID
BRCA 1.25× (4096 × 4096)	57k	Graikos *et al.* [10]	976k patches of 1024 × 1024	2.75	**11.30**
		∞-Brush	65536 pixels of 57k full-size images	**2.63**	14.76
		∞-Brush ✗ Cross-attention neural operator		3.81	16.28

Experiments – Controllable Large Image Generation



Large images (1024×1024) generated from ∞-Brush, and the corresponding reference real images used to generate them.

Experiments – Controllable Large Image Generation



Large images (2048×2048 and 1024×1024) generated from ∞-Brush 🎨 , and the corresponding reference real images used to generate them.

Experiments – Controllable Large Image Generation

Table 3: Performance on controllable large image synthesis on BRCA 5× and NAIP dataset at 1024×1024 resolution. ∞-Brush outperforms other methods in global structure accuracy, with a marginal trade-off in fine detail as reflected in Crop FID.

Dataset	# Images	Method	Training Config.	CLIP FID	Crop FID
BRCA 5× (1024 × 1024)	976k	SDXL [25]	976k full-size images	6.64	**6.98**
		Graikos *et al.* [10]	15M patches of 256 × 256	7.43	15.51
		∞-Brush	256 × 256 pixels of 976k full-size images	**3.74**	17.87
NAIP (1024 × 1024)	35k	SDXL [25]	35k full-size images	10.90	**11.50**
		Graikos *et al.* [10]	667k patches of 256 × 256	6.86	43.76
		∞-Brush	256 × 256 pixels of 35k full-size images	**6.32**	48.65

Experiments – Computing Resource Evaluation

Table 4: Computing resources requirements for different diffusion models. our ∞-**Brush** maintains a constant parameter count and batch size across resolutions, highlighting its efficiency and scalability for controllable large image generation.

Method	# Params.	Training at 1024×1024		Training at 4096×4096	
		Max batch size	Epoch time	Max batch size	Epoch time
SDXL [25]	3.5B	4	140 hr	O.O.M	1000 hr (estimated) currently infeasible
Graikos *et al.* [10]	860M	100	45 hr	4	300 hr
∞-**Brush**	450M	20	12 hr	20	12 hr

Thanks!