Learned representation-guided diffusion models for large-image generation

Alexandros Graikos* Srikar Yellapragada* Minh-Quan Le
Saarthak Kapse Prateek Prasanna Joel Saltz Dimitris Samaras
Stony Brook University

Abstract

To synthesize high-fidelity samples, diffusion models typically require auxiliary data to guide the generation process. However, it is impractical to procure the painstaking patch-level annotation effort required in specialized domains like histopathology and satellite imagery; it is often performed by domain experts and involves hundreds of millions of patches. Modern-day self-supervised learning (SSL) representations encode rich semantic and visual information. In this paper, we posit that such representations are expressive enough to act as proxies to fine-grained human labels. We introduce a novel approach that trains diffusion models conditioned on embeddings from SSL. Our diffusion models successfully project these features back to high-quality histopathology and remote sensing images. In addition, we construct larger images by assembling spatially consistent patches inferred from SSL embeddings, preserving long-range dependencies. Augmenting real data by generating variations of real images improves downstream classifier accuracy for patch-level and larger, image-scale classification tasks. Our models are effective even on datasets not encountered during training, demonstrating their robustness and generalizability. Generating images from learned embeddings is agnostic to the source of the embeddings. The SSL embeddings used to generate a large image can either be extracted from a reference image, or sampled from an auxiliary model conditioned on any related modality (e.g. class labels, text, genomic data). As proof of concept, we introduce the text-to-large image synthesis paradigm where we successfully synthesize large pathology and satellite images out of text descriptions.\footnote{Equal contribution. Correspondence to agraikos@cs.stonybrook.edu}

1. Introduction

Diffusion models produce high-quality and diverse samples across a spectrum of generative tasks \cite{8, 24}. This leap forward has been enabled by the simultaneous curation of large-scale multi-modal datasets \cite{40} and the development of efficient conditioning mechanisms \cite{35, 37}. The key to unlocking the models’ capabilities is to integrate auxiliary information during training and inference \cite{8, 15, 28}.

Large-scale human-annotated datasets are mostly limited to image-caption pairs, collected from easily accessible online repositories and labeled by non-expert annotators. However, in domains such as digital histopathology and remote sensing, where gigapixel scale images provide vast amounts of unlabeled data, annotation proves challenging. Moreover, the process requires expert knowledge and is more difficult at finer scales, i.e., captioning large crops of the gigapixel image is simpler than captioning smaller patches. Based on our estimates (see supplemental), annotating the entire TCGA-BRCA dataset with captions would take \( \approx 40,000 \) hours of pathologist’s time. Replicating the impressive results of diffusion models in these domains has been limited by the scarcity of fine-grained per-image conditioning, vital for high-quality image synthesis \cite{29}.

Modern-day self-supervised learning (SSL) representations \cite{5} encode rich semantic and visual information. Features from trained self-supervised models serve as compact image representations and are widely used to perform discriminative downstream tasks successfully \cite{6, 11, 45}, proving that these compressed representations indeed encode...
useful semantic information about the images. We hypothesize that such SSL representations are already expressive enough to act as proxies to fine-grained human labels. If this is true, these representations should be able to condition the training of effective diffusion models in these domains. In this novel approach, we utilize self-supervised feature extractors as image annotators; these features provide necessary per-image conditioning signals at the highest resolutions, required for diffusion model training.

Our experiments show that conditioning with expressive self-supervised features leads to precise control over the image content. The SSL features are adept at identifying complex patterns and structures in the images, while the diffusion model learns to translate them into visual components accurately (Fig. 3). This motivates us to perform large-image synthesis by locally controlling the appearance using SSL conditioning and dictating the global structure through the spatial arrangement of the conditions.

Our approach synthesizes large images in a patch-based manner, using a single image diffusion model at the highest resolution. We represent a large image as a grid of SSL embeddings, where each serves as a representation of a large image neighborhood. The whole image is then synthesized by generating consistent patches that capture both the local properties, as given by the local patch conditioning, and the spatial arrangement of the conditioning features. If we change this spatial arrangement, we are, in fact, editing how semantic elements are arranged globally in the large image. This strategy enables the controllable generation of images of virtually any size without significantly increased computation compared to the base patch-level model.

To generate a large image, our approach requires the patch diffusion model and the spatially-arranged conditioning. We can start with a reference large image as source and extract SSL embeddings from non-overlapping segments, enabling our method to synthesize a variation of the original image (Fig. 1).

Utilizing SSL embeddings as conditions allows us to have the necessary control over image generation, at the expense of an explainable and easy-to-use conditioning mechanism. Nevertheless, we argue that since generating images from learned embeddings is agnostic to the embedding source, there are simple ways to combine control over generated images with explainability. We propose training auxiliary models to transform higher-level conditioning signals, such as text captions, to the learned patch representations. To demonstrate this versatility, we introduce text-to-large-image synthesis by training an auxiliary model to sample a spatial arrangement of embeddings from a text description.

We train patch-level diffusion models using self-supervised features as conditioning on digital histopathology (TCGA [4]) and satellite image (NAIP [43]) datasets. We perform extensive evaluations and demonstrate the advantages of SSL conditioning and our large-image generation framework on synthesis and classification tasks. Our model achieves exceptional patch-level and large-image quality, the ability to improve classifiers through data augmentation even when synthesizing out-of-distribution data, and effective fusion of diffusion and SSL features for downstream applications. Finally, we are the first to perform text-to-large image synthesis, which should be of significant community interest as vision-language models (VLMs) for pathology and satellite images gain traction.

In summary, our contributions are as follows:

- We develop a novel method to train diffusion models with self-supervised learning features as conditioning and generate high-quality images in the histopathology and satellite image domains.
- We present a framework for large-image synthesis, based on self-supervised guided diffusion, that maintains contextual integrity and image realism over large areas.
- We demonstrate the applicability of our model in various classification tasks and showcase its unique ability to augment out-of-distribution datasets.
- We introduce text-to-large image generation for digital histopathology and satellite images, highlighting the versatility of our approach.

2. Related work

Diffusion models: Introduced for image generation by Ho et al. [16], diffusion models have evolved considerably. These enhancements include class conditioning [28], architectural improvements and gradient-based guidance [8], and classifier-free guidance [15]. Latent Diffusion Models (LDMs) [37] proposed a two-step training process with a Variational Autoencoder (VAE) compressing input images into a lower-dimensional latent space and a diffusion model trained in this latent space. Denoising Diffusion Implicit Models (DDIM) [42] accelerate the sampling process by $10 − 50 \times$. Self-guided diffusion models [17] also utilize self-supervised learning by quantizing SSL embeddings. In contrast to our approach, their quantization discards useful information from the SSL embeddings.

Training generative models directly at the gigapixel resolution is infeasible. An alternative is to generate the images in a coarse-to-fine manner hierarchically. This has already been applied to natural images, where chaining multiple diffusion models generates images up to $1024 \times 1024$ resolution [32, 39]. However, it is still limited, as it substantially increases the parameter count and is inherently constrained by the final target resolution.

In the context of digital histopathology, works have been limited to training unconditional [26] or class-conditioned [27, 47] diffusion models. For text conditioning, pathology text reports were used to provide context on the whole-slide scale [48]. Similar approaches have been
Figure 2. (a) We train diffusion models on patches $I$ (e.g. the one in the green box) taken from a large image conditioned on SSL embeddings. (b) We present our large image generation framework in 4 steps: (i) We extract a set of spatially arranged embeddings from a reference image or sample them from an auxiliary model. (ii) For every location $(i, j)$, we compute a conditioning vector $\lambda_{i,j}$ by interpolating the spatial grid of embeddings. (iii) At every diffusion step, we denoise the patch $F(i, j)$ using the conditioning $\lambda_{i,j}$. (iv) The next step is computed by averaging the denoising updates of all patches that overlap at $(i, j)$.

applied to satellite data \[10, 41\]. Apart from high-quality images without manual annotation, our SSL conditioning is also necessary for large-image synthesis, as all previous conditioning methods would not be sensitive to the intricate differences between neighboring patches. In recent work, DiffInfinite \[1\] explored large-image generation using segmentation masks as conditioning. We argue this is still sub-optimal as it requires accurate, human-annotated masks for training and a mask-generating model during inference.

**Self-Supervised Learning** Self-supervised learning (SSL) refers to both discriminative \[2, 13\] and distillation \[5\] approaches that aim to learn representations of the data without supervision. In this work, we are mainly interested in self-supervised learning for histopathology. Notable recent developments include the Hierarchical Image Pyramid Transformer (HIPT) \[6\], which utilizes the inherent hierarchical structure of Whole Slide Images (WSIs), CTransPath \[45\], a hybrid CNN and multi-scale Swin Transformer model, and iBOT \[11\], a masked image modeling method. These models excel as patch-level feature extractors for WSIs by leveraging the unique structure of the large-image data and capture the important semantic information that we require for conditioning a generative model.

### 3. Method

We propose training a diffusion model on patches $I$ drawn from large histopathology and satellite images, using self-supervised embeddings $y$ as conditioning. Furthermore, we present a patch-based approach that utilizes the SSL-conditioned diffusion model to synthesize arbitrarily large images. An overview of our method is presented in Fig. 2.

#### 3.1. Learned representation-guided diffusion

Given a pre-trained self-supervised feature extractor, we employ LDMs \[37\] to learn the distribution over the large-image patches $p(I)$. LDMs are comprised of three components: an image-compressing Variational Autoencoder (VAE) that transforms the input images to latent representations, a U-Net denoiser to learn a denoising diffusion process that transforms Gaussian noise to latents, and a conditioning mechanism that controls the diffusion process. The conditioning is performed in our setting with a single vector $y$, obtained from the self-supervised model, integrated via a cross-attention mechanism. We train the LDM using pairs of image patches $I$ and the corresponding extracted self-supervised embeddings $y$.

#### 3.2. Large image generation

Our goal is to synthesize high-quality large images, that not only capture global structure but also maintain spatial consistency. As shown in Fig. 2, we replicate the semantics in each patch with SSL-guided LDM. At the same time, we preserve the global arrangement of these semantics as defined by the grid of patches in the reference image. Future
work can explore alternative approaches to ensure spatial alignment, including using topological constraints or integrating low-resolution information.

We follow the MultiDiffusion [3] methodology to generate large images using only a patch-based diffusion model. We can represent the diffusion model as a learned mapping from images and conditions to images:

$$\Phi: \mathcal{I} \times \mathcal{Y} \rightarrow \mathcal{I},$$  
(1)

where \( \mathcal{I} = \mathbb{R}^{H \times W \times C} \) are the large-image patches and \( \mathcal{Y} = \mathbb{R}^d \) are the per-patch conditions. To generate a patch, we initialize \( I_0 \sim \mathcal{N}(0,1) \) and sequentially transform to a “clean” image \( I_0 \) following the trained LDM:

$$I_{t-1} = \Phi(I_t \mid y)$$  
(2)

for \( t = T, T-1, \ldots, 1 \). We assume that a large image can be formed as a spatial grid of \( P \times P \) patches of size \( H \times W \). We then define the large-image diffusion model as

$$\Psi: \mathcal{J} \times \mathcal{Z} \rightarrow \mathcal{J}$$  
(3)

where \( \mathcal{J} = \mathbb{R}^{PH \times PW \times C} \) are the large images and \( \mathcal{Z} = \mathcal{Y} \times \mathcal{Y} \times \cdots \times \mathcal{Y} \) are the conditioning vectors of all patches.

The process \( \Psi(J_t \mid z) = J_{t-1} \), with \( J_T \sim \mathcal{N}(0,1) \), can be approximated by first defining mappings between the two different image and conditioning spaces:

$$F_{i,j}^{\text{proj}}: \mathcal{J} \rightarrow I, \quad \lambda_{i,j}: \mathcal{Z} \rightarrow \mathcal{Y}$$  
(4)

where, for the case of large-image generation, we set \( F_{i,j}^{\text{proj}} \) to be a crop (projection) of the large image, centered at \( i, j \), and \( \lambda_{i,j} \) the conditioning \( y_{i,j} \in \mathbb{R}^d \) at \( i, j \). Since we only have conditioning vectors at the centers of the patches, we can use spatial interpolation algorithms to implement \( \lambda \). This assumes that the interpolant \( y_{i,j} \) is a valid conditioning vector for the diffusion process, which we validate experimentally. Then, \( \Psi \) is defined as

$$\Psi(J_t \mid z) = \arg \min_{J \in \mathcal{J}} \sum_{i,j} \| F_{i,j}^{\text{proj}}(J) - \Phi(F_{i,j}^{\text{proj}}(J_t) \mid \lambda_{i,j}(z)) \|^2 $$  
(5)

which can be solved in closed-form by setting each pixel \( i, j \) of \( J \) to the average of all the patch-diffusion updates

$$\Psi(J_t \mid z) = \sum_{i, j} F_{i,j}^{\text{anproj}}(1) \sum_{k,l} F_{k,l}^{\text{anproj}}(1) \odot F_{i,j}^{\text{anproj}}(\Phi(F_{i,j}^{\text{proj}}(J_t) \mid \lambda_{i,j}(z)))$$  
(6)

where \( 1 \) denotes a patch image where all values are set to 1 and \( F_{i,j}^{\text{anproj}} \) is the inverse mapping of pixels from the crop centered at \( i, j \) back to the large image.

We are able to generate images larger than the ones produced by the patch diffusion model. At the same time, we control what each patch looks like, which is crucial in maintaining the semantic integrity of the larger image. The self-supervised conditions can capture the variations between neighboring patches necessary for producing realistic results. Generating large images with coarser conditioning, such as global text prompts [48], would lead to uniform texture regions (see supplementary).

### 3.3. Controllable large-image synthesis

Although the grid of self-supervised embeddings \( z \) cannot be manipulated in a human-interpretable manner, we argue that it is simple to assert more control over the generated images. As illustrated in Fig. 2, this control can be attained by training auxiliary models \( p(z \mid c) \) that translate higher-level conditioning signals, such as text captions \( c \), to learned patch representations \( z \).

Since there is no available dataset of paired large images and captions, we resort to pre-trained multi-modal models, such as Quilt [18], CLIP [34] and BLIP [22], to provide the text conditioning. We construct training sets by extracting self-supervised embeddings from training-set large images and pairing them with multi-modal image embeddings or generated captions. During inference, we first sample SSL embeddings from the learned distribution \( p(z \mid c) \), then utilize our patch diffusion models to synthesize a large image.

### 4. Image Generation Experiments

#### 4.1. Datasets

We train diffusion models on digital histopathology images from The Genome Cancer Atlas (TCGA) [4] and satellite imagery from the National Agriculture Imagery Program (NAIP) [43]. Specifically, we used the TCGA-BRCA (Breast Invasive Carcinoma Collection), TCGA-CRC (COAD + READ Colorectal Carcinoma) datasets, and the Chesapeake Land Cover dataset [36].

For the TCGA-BRCA and TCGA-CRC datasets, we use images at \( 20 \times \) magnification, and developed diffusion models conditioned on embeddings from HIPT [6] and iBOT [11], which were pre-trained on PanCancer TCGA. In the case of the HIPT model, we specifically used its patch-level ViT. Additionally, for the TCGA-BRCA dataset, we train a model at \( 5 \times \) magnification using embeddings from CTransPath [44] for additional evaluations.

The Chesapeake Land Cover dataset dataset contains 732 NAIP tiles, each measuring a 6km \( \times \) 7.5 km area at 1m resolution. We extract \( 256 \times 256 \) non-overlapping pixel patches, resulting in 667,000 patches total. Given the absence of publicly available self-supervised learning models tailored to the NAIP data, we train a Vision Transformer (ViT-B/16) [9] using the DINO framework [5]. We then use the learned DINO embeddings to train the diffusion model on pairs of image patches and self-supervised embeddings.
For patch-level augmentation, we employ the NCT-CRC dataset [19], which has 100,000 Colorectal cancer (CRC) patches from 86 patients. Each 224 × 224 pixel patch at 20× magnification is annotated with one of nine distinct tissue class labels.

For large-image augmentation, apart from TCGA-BRCA, we also use the BACH dataset [33]. Introduced in the ICIAR 2018 Grand Challenge, the dataset contains 400 H&E-stained Breast Cancer images of size 2048 × 1536 pixels, evenly distributed across four categories: normal, benign, in-situ carcinoma, and invasive carcinoma.

4.2. Implementation details

For all our experiments, we train the LDM on 256 × 256 pixel patches, following PathLDM [48], which fine-tunes an ImageNet-trained [38] U-Net denoiser and uses a 4× downsampling VQ-VAE, instead of the default LDM configuration. These modifications were deemed necessary for applying LDMs on large-image domains.

We train our models on 6 NVIDIA RTX 8000 GPUs, with a batch size of 100 per GPU, utilizing code and pre-trained checkpoints from LDM [37]. We set the learning rate at 10^{-4} with a warmup of 10,000 steps. We apply DDIM with 50 steps and a guidance scale of 1.75 for both patch sampling and large-image generation. When generating large images we apply the patch diffusion model with a stride > 1 depending on the desired target quality.

4.3. Image quality results

In Fig. 3, we present synthetic patches and large images from our TCGA-BRCA and NAIP models, along with the corresponding references from which the self-supervised embeddings were extracted. We evaluate our method’s per-patch and large-image generation quality by computing FID scores [14] using the Clean-FID implementation [31]. We generate 10,000 patches (256 × 256) and 3,000 large images (1024 × 1024) from diffusion models trained on the TCGA-BRCA, TCGA-CRC, and NAIP datasets. Since our generative model requires a self-supervised conditioning vector y (or multiple vectors z) for each synthetic image (or large image), we randomly sample embeddings from reference images in the training set to generate images for evaluation.

For patches, we measure FID against the real images (“Vanilla FID”). For large images, we follow the evaluation strategy of MultiDiffusion [3] and use FID to compare the distribution of 256 × 256 crops from synthesized large images to that of real image patches of the same size (“Crop FID”). We also measure FID directly between the large images and ground truth data using CLIP [34] (“CLIP FID”).

To evaluate the similarity between the reference and generated large images, we resize from 1024 × 1024 to 256 × 256 and use an SSL model to extract embeddings. We compute the cosine similarity between the paired embeddings (“Embedding Similarity”).

Patch-level quality: Our models achieve low patch FID scores across all datasets (Tab. 1). For TCGA-BRCA patches, our “Vanilla FID” of 6.98 is comparable to the current state-of-the-art [48] (7.64 at 10×). We attain similar, low FID scores for the smaller CRC and NAIP datasets.

Synthetic large image quality: We present the “Crop FID”, “CLIP FID” and “Embedding Similarity” results for large images in Table 1. The “Crop FID” of the large im-
As shown in Fig. 5, Image Augmentation Experiments

tiveness in preserving the integrity of context and key features of real images for training purposes.

We evaluate the contextual similarity between synthetic and reference large images for the BRCA and CRC data. We compute cosine similarity between large images using CTransPath embeddings. Our BRCA and CRC models demonstrate high “Embedding Similarity” scores of 0.924 and 0.938, respectively, reflecting our framework’s effectiveness in preserving the integrity of context and key features on the large-image scale.

5. Image Augmentation Experiments

As shown in Fig. 3, apart from visual fidelity, the synthetic images preserve the intricate characteristics of the reference images, both on the patch and large-image scale. These characteristics include the nuanced textural elements, histological staining, and cell structure. This correspondence between real and synthetic images demonstrates the detailed and varied information captured by the self-supervised embeddings used as conditioning. In conjunction with the high “Embedding Similarity”, it justifies using variations of images generated from SSL embeddings for patch and large-image level data augmentation.

Having a powerful generative model enables us to perform data augmentation for tasks where we can control the augmented image label using conditioning. In our setting, our diffusion models are not trained with class labels; instead, we synthesize a novel image using the conditioning from a reference patch or region.

We assume that i) the self-supervised embedding used as conditioning contains information about the target label and ii) the diffusion-generated variations of an image do not alter this target label information. We validate both assumptions experimentally, by augmenting training sets on patch and large-image level tissue classification tasks, including a Multiple Instance Learning (MIL) task.

Large-image augmentation on TCGA-BRCA: We examine two histopathology slide-level binary classification tasks on TCGA-BRCA: BRCA Subtyping (Invasive Ductal Carcinoma (IDC) vs Invasive Lobular Carcinoma (ILC)) and HRD prediction. We utilize a minimal dataset, just 10% of real WSI data (100 WSIs), to train MIL algorithms. We generate an equal set of synthetic images for 100 additional WSIs using training set images as reference.

We employ 10-fold cross-validation to divide our dataset into training and testing segments. Within each fold, two multiple instance learning (MIL) models, CLAM-SB [25] and DSMIL [21], are trained on two sets: one with real data and another combining real and synthetic data. To train the MIL models, we extract features using the CTransPath ViT [44]. The results, detailed in Table 2, indicate that models trained on the augmented datasets consistently surpass their real-only counterparts, regardless of the MIL algorithm used. This demonstrates the value of the synthetic images generated by our method, confirming their efficacy as comparable to real images for training purposes.

Large-image augmentation on BACH: We double the training set of BACH [33] by adding as many synthetic large-images, produced by the TCGA-BRCA diffusion model. From each 2048 × 1536 pixel training set image, we extract a 8 × 6 SSL embedding grid to generate a variation with the same label. For classification, we employ a ConvNeXt V2_huge [46] model pre-trained on ImageNet. We train a 2-layer MLP classifier on top of the penultimate layer features and evaluate it on the official test set.

The results, presented in Table 3a, reveal a notable improvement in classifier performance, from 78% to 83%. This improvement again confirms the high quality of our synthetic data while also highlighting the versatility of our approach. Despite being trained on 256 × 256 patches from TCGA-BRCA, the model generalizes to produce realistic large images from a completely different dataset. We attribute this generalization capability to the expressiveness of the SSL features and the potency of the diffusion model.
to accurately portray them in images. While our model does not reach the current SoTA accuracy of 87%, achieved by an ensemble approach [7], the simplicity and ability to integrate with other models make for a noteworthy contribution.

<table>
<thead>
<tr>
<th>Method</th>
<th>1% Real + synthetic</th>
<th>10% Real + synthetic</th>
<th>20% Real + synthetic</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLAM-SB</td>
<td>0.725</td>
<td>0.812</td>
<td>0.898</td>
</tr>
<tr>
<td>DSML</td>
<td>0.609</td>
<td>0.659</td>
<td>0.856</td>
</tr>
</tbody>
</table>

Table 2. The inclusion of synthetic data consistently enhances AUC across various MIL architectures and BRCA tasks. The dataset contains 1000 real images, so “10% + synthetic” indicates training with 100 real and 100 synthetic WSIs, with the remainder used for testing.

Table 3. Our data augmentations provide notable improvements for the BACH (a) and CRC-VAL-HE (b) datasets. Notably, the diffusion training data does not overlap with the data of the augmented datasets.

**Patch-level image augmentation:** We further investigate our out-of-distribution generalization capabilities by augmenting the NCT-CRC dataset [19]. We leverage our diffusion model trained on the TCGA-CRC data, which does not overlap with NCT-CRC, conditioned on embeddings from an iBOT ViT [11]. We generate an augmented dataset of equal size to the original by using the SSL embedding of every patch in the NCT-CRC training set to synthesize a corresponding image of the same label.

We train an ImageNet pre-trained ResNet-50 [12] network on three splits: real images only, synthetic images only, and a combination of both. Evaluation on the CRC-VAL-HE-7K test set demonstrates a significant performance increase when synthetic data is introduced, with classifier accuracy rising from 93.8% to 96.27% as presented in Table 3b. We match the current SoTA [20], which used an ensemble of deep models, with a model agnostic approach. As in the previous experiment, the diffusion model, now trained on the significantly smaller TCGA-CRC data, can also effectively synthesize images from a completely different dataset by only controlling the self-supervised conditioning.

6. **Text-to-large image synthesis**

For previous tasks, our large image generation approach synthesizes variations of an existing set of images from the pre-computed self-supervised embeddings. In Sec. 3.3 we discussed how to control large image generation with auxiliary signals from any domain (class labels, text captions, etc.), by training models $p(z | c)$ that can be combined with the embedding-conditioned image generation. We demonstrate controllable image synthesis with text-to-large image generation experiments on histopathology and satellite data. We measure the similarity between synthetic images and the text prompts used in generating them using vision-language models (VLMs).

**Text-to-large histopathology images:** We utilize the CRC and BRCA diffusion models to generate 1024 × 1024 pixel images from text prompts. We first construct training sets by pairing 4 × 4 SSL embedding grids $z$ from large BRCA and CRC images, with their corresponding Quilt [18] image embeddings $c$. We then train an auxiliary diffusion model to sample $p(z | c)$. During inference, we use a text embedding $c'$ as a proxy for the image embedding, to sample $z$ and synthesize a large image. To bridge the gap between the image and text Quilt embeddings [23, 30] we perturb the image embeddings when training the auxiliary diffusion model with Gaussian noise of variance $\sigma^2 = 0.1$.

To evaluate the text-to-large image pipeline, we generate images from a pre-defined set of classes described in natural language; non-malignant benign tissue, malignant in-situ carcinoma, malignant invasive carcinoma, normal breast tissue for BRCA and colon adenocarcinoma, benign colonic tissue for CRC. We construct the confusion matrix of zero-shot classifiers on the synthesized data. We used two different VLMs as zero-shot classifiers, Quilt and BiomedCLIP [49]. The results presented in Fig. 4 demonstrate our ability to synthesize images consistent with the text prompts. The capabilities of the VLM used, limit our synthetic image generation. The lower performance of BRCA vs. CRC is consistent with the results reported in Quilt [18]. Furthermore, we asked an expert pathologist to classify 100 synthetic CRC images as benign or adenocarcinoma images. Their evaluation showed an 89.9% agreement rate with the labels used for image generation, indicating consistency in our text-to-large image pipeline. We show examples of synthesized images in the supplementary.

**Text-to-large satellite images:** To synthesize novel satellite images we first create a training set of 30k 1024 × 1024 pixel large NAIP images, and pair them with captions
from a BLIP model [22]. We train a diffusion model to sample the $4 \times 4$ SSL embeddings $z$ from the captions $c$. To evaluate, we create a separate set of 1000 NAIP image-caption pairs and measure the CLIP similarity between generated images and the given captions. We achieve a CLIP similarity score of 0.22, showing that we can effectively learn the mapping from text to large images with this hierarchical approach. Although our CLIP similarity is slightly worse than the scores reported for text-to-image Stable Diffusion models ($>0.24$) [32], we expect this drop in performance as we trained with machine-generated captions. Training and generated image-caption pairs are provided in supplementary.

We further evaluate our patch-generating models by posing the question does the diffusion model learn more about the data than the self-supervised learning model? We hypothesize that by combining the pre-trained self-supervised embedding with the denoising task we can improve the learned representations of the data, leading to better performance in downstream tasks, which are performed with features from self-supervised learning. To validate this hypothesis, we utilize the trained diffusion model as a feature extractor and apply a Multiple-instance learning (MIL) approach for the slide-level classification task of subtyping Breast Cancer. For each patch in a whole-slide image (WSI) we first extract the self-supervised embeddings, which are then used as conditioning to obtain features from the denoiser’s U-Net bottleneck layer at a fixed timestep $t = 50$.

In Tab. 4, we evaluate the effectiveness of our novel fusion of generative diffusion and self-supervised embeddings. We compare the performance of MIL algorithms [21, 25] using features derived from our approach at $20 \times$ and $5 \times$ magnification (LDM$_\text{HIPT}$/CTransPath) against using only the self-supervised conditioning (HIPT/ CTransPath). We used a 10-fold cross-validation strategy consistent with the data splits from HIPT [6], training the MIL algorithms on both the full dataset (100%) and a reduced subset (25%).

The results indicate that integrating self-supervised features into the diffusion model as conditioning leads to learning better representations and improves whole-slide classification. By fusing the generative knowledge from the diffusion process with the discriminative capabilities of the self-supervised embeddings, we construct a successful model for both discriminative and generative tasks.

Table 4. 10-fold cross-validation AUC for BRCA Histological subtyping. LDM$_\text{HIPT}$ denotes diffusion features conditioned on HIPT embeddings. The fusion of SSL and diffusion features outperforms the SSL features by themselves.

<table>
<thead>
<tr>
<th>Mag</th>
<th>Features</th>
<th>25% training</th>
<th>100% training</th>
</tr>
</thead>
<tbody>
<tr>
<td>20×</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>HIPT</td>
<td>0.788</td>
<td>0.861</td>
</tr>
<tr>
<td></td>
<td>LDM$_\text{HIPT}$</td>
<td>0.842</td>
<td>0.894</td>
</tr>
<tr>
<td>5×</td>
<td>CTransPath</td>
<td>0.900</td>
<td>0.919</td>
</tr>
<tr>
<td></td>
<td>LDM$_\text{CTransPath}$</td>
<td>0.913</td>
<td>0.923</td>
</tr>
</tbody>
</table>

7. Combining self-supervised embeddings with diffusion

We presented a novel approach to training diffusion models in large-image domains, such as digital histopathology and remote sensing. We overcome the need for fine-grained annotation by introducing self-supervised representation guided diffusion models, achieving remarkable image synthesis results on the patch level. Our approach also enables us to synthesize high-quality large images, where we have the ability to dictate the global structure by controlling the spatial arrangement of the conditions. We evaluated the usefulness of our synthetic images on a number of patch and large image-level tasks, as well as introduced a text-to-large image generation framework. Naively augmenting whole-slide images is a time-consuming process. We leave to future work the exploration of adaptive augmentation strategies that choose which image parts to augment. We believe these results illustrate the great potential for this technology to lead to bespoke foundational models for specialized domains, comparable to existing models for natural images.

8. Conclusion

Acknowledgements This research was partially supported by NCI awards SU24CA215109, 1R21CA258493-01A1, UH3CA225021, NSF grants IIS-2123920, IIS-2212046 and Stony Brook Profund 2022 seed funding. We thank Rajarsi Gupta for his valuable feedback.
References


[49] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Navyen