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. [1]

1. Introduction

Diffusion models produce high-quality and diverse samples across a spectrum of generative tasks [8, 24]. This leap forward has been enabled by the simultaneous curation of large-scale multi-modal datasets [41] and the development of efficient conditioning mechanisms [36, 38]. The key to unlocking the models' capabilities is to integrate auxiliary information during training and inference [8, 15, 29].

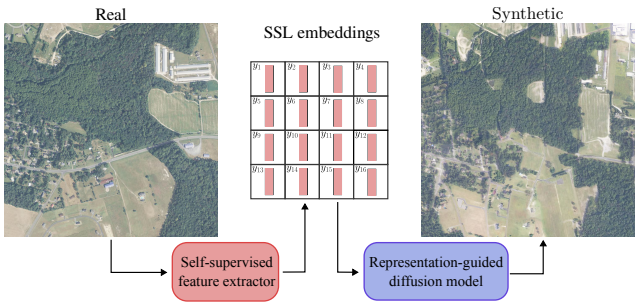

Figure 1. We propose using SSL features to condition diffusion models. This allows us to construct large images by assembling consistent patches inferred from a spatial arrangement of SSL embeddings. The generated image retains the semantics of the embeddings used as a condition, maintaining the forested and open areas from the reference. Best viewed zoomed-in.

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 a finer scale, 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 ≈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 [30].

Modern-day self-supervised learning (SSL) representations [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 [6, 11, 46], proving that these compressed representations indeed encode

*Equal contribution. Correspondence to agraikos@cs.stonybrook.edu
[1] Code is available at this link

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 [44]) 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 [29], architectural improvements and gradient-based guidance [8], and classifier-free guidance [15]. Latent Diffusion Models (LDMs) [38] 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) [43] 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 1024×1024 resolution [33, 40]. 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 [27] or class-conditioned [28, 48] diffusion models. For text conditioning, pathology text reports were used to provide context on the whole-slide scale [49]. Similar approaches have been

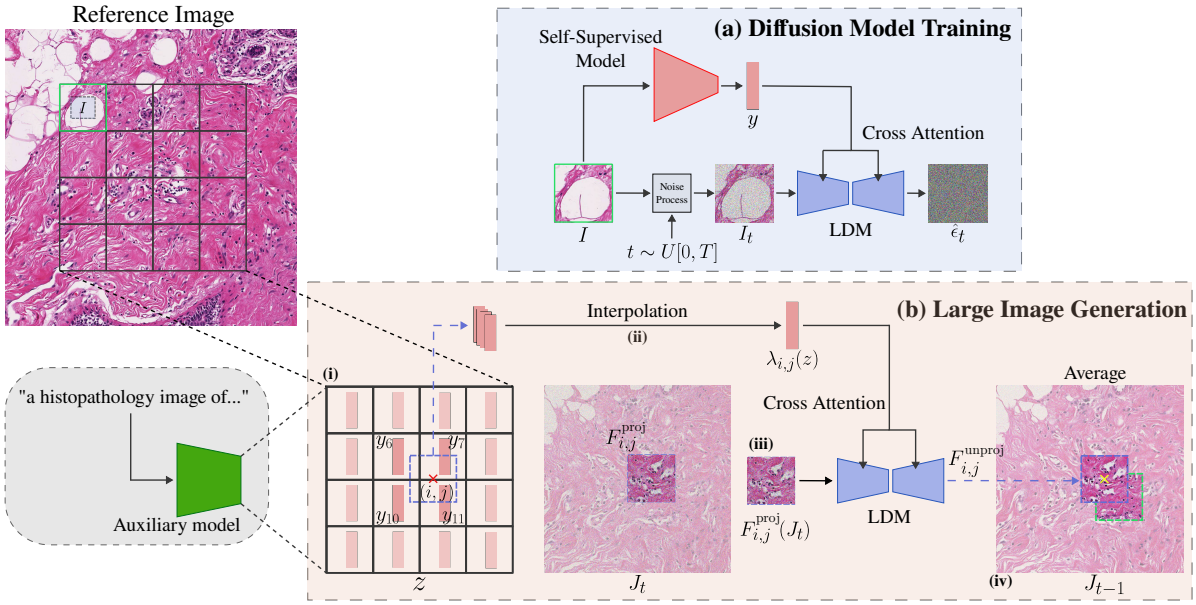

Figure 2. (a) We train diffusion models on patches I (e.g. the one in the green box) taken from a large image conditioned on SSL embeddings. (b) We present our large image generation framework in 4 steps: (i) We extract a set of spatially arranged embeddings from a reference image or sample them from 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, 42]. 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 suboptimal 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 [46], 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 [38] to learn the distribution over the large-image patches $p(\mathcal{I})$. 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:

$$\Phi : \mathcal{I} \times \mathcal{Y} \to \mathcal{I}, \tag{1}$$

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 $I_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and sequentially transform to a "clean" image I_0 following the trained LDM:

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

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

$$\Psi : \mathcal{J} \times \mathcal{Z} \to \mathcal{J} \tag{3}$$

where $\mathcal{J} = \mathbb{R}^{PH \times PW \times C}$ are the large images and $\mathcal{Z} = \underbrace{\mathcal{Y} \times \mathcal{Y} \cdots \times \mathcal{Y}}_{P^2}$ are the conditioning vectors of all patches. The process $\Psi(J_t \mid z) = J_{t-1}$, with $J_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, can be approximated by first defining mappings between the two different image and conditioning spaces:

$$F^{\text{proj}}_{i,j} : \mathcal{J} \to I, \quad \lambda_{i,j} : \mathcal{Z} \to \mathcal{Y} \tag{4}$$

where, for the case of large-image generation, we set $F^{\text{proj}}_{i,j}$ 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 λ. This assumes that the interpolant $y_{i,j}$ is a valid conditioning vector for the diffusion process, which we validate experimentally. Then, Ψ is defined as

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

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

$$\Psi(J_t \mid z) = \sum_{i,j} \frac{F^{\text{unproj}}_{i,j}(\mathbf{1})}{\sum_{k,l} F^{\text{unproj}}_{k,l}(\mathbf{1})} \otimes F^{\text{unproj}}_{i,j}(\Phi(F^{\text{proj}}_{i,j}(J_t) \mid \lambda_{i,j}(z))) \tag{6}$$

where $\mathbf{1}$ denotes a patch image where all values are set to 1 and $F^{\text{unproj}}_{i,j}$ 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 [49], 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 [35] 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) [44]. Specifically, we used the TCGA-BRCA (Breast Invasive Carcinoma Collection), TCGA-CRC (COAD + READ Colorectal Carcinoma) datasets, and the Chesapeake Land Cover dataset [37].

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 [45] 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 256×256 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.

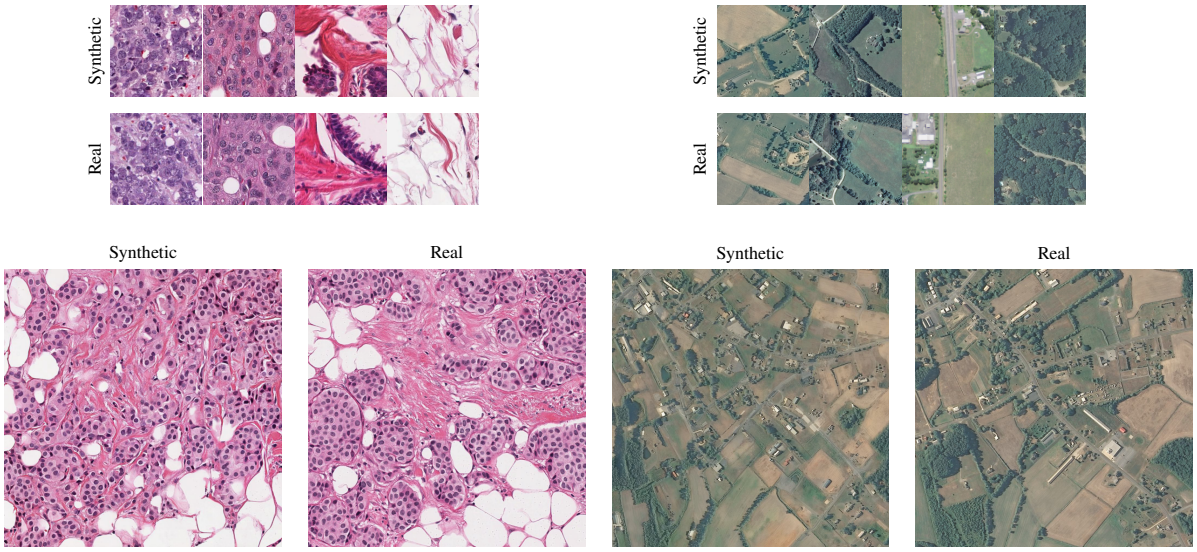

Figure 3. (Top) Patches (256 × 256) from our models, and the corresponding reference real patches used to generate them. The SSL-guided LDM replicates the semantics of the reference patch. (Bottom) Large images (1024 × 1024) from our models, and the corresponding reference real images used to generate them. We preserve the global arrangement of the semantics defined in the reference image.

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 [34]. 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 [49], which fine-tunes an ImageNet-trained [39] 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 [38]. 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 [32]. 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 [35] ("**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 [49] (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-

Dataset	# Training images	Patch level	Large image level		
		Vanilla FID	Crop FID	CLIP FID	Emb similarity
BRCA 20x	15M	6.98	15.51	7.43	0.924
CRC 20x	8M	6.78	8.8	7.34	0.938
NAIP	667k	11.5	43.76	6.86	-
BRCA 5x	976k	9.74	-	6.64	-

Table 1. FID scores for our generated patch and large images. Our patch-level BRCA model is on par with SoTA [49] (7.64 at 10 $\times$). "CLIP FID" and "Embedding Similarity" demonstrate our large images' realism and contextual accuracy.

ages is comparable to the "Vanilla FID" on the BRCA and CRC datasets, showing that patches from synthesized large images have similar semantic content to the ground truth patches. We attribute the worse "Crop FID" for the NAIP model to the limited number of samples available for both SSL and diffusion model training. The "Crop FID" is consistently higher than the patch-level FID, which is expected as the large-image generation framework only approximates the distribution of the large images and does not have access to the true conditioning at every location.

Our low "CLIP FID" scores indicate that the generated large images are similar to real images when resized to 224×224 pixels. This indicates that our SSL-guided large-image generation successfully retains the larger-scale semantic arrangements of real data. For NAIP, our model is not as good at synthesizing high-frequency details, which explains the large discrepancy between "Crop FID" and "CLIP FID". Additionally, when comparing the "CLIP FID" of large images synthesized by the $20\times$ BRCA model and resized to $5\times$, to images from a model trained directly on $5\times$ data, we see minimal difference (7.43 vs. 6.64).

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 [26] 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 [45]. 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 [34] 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 [47] 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 apporach. 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

Method	TCGA-BRCA Subtyping					TCGA-BRCA HRD				
	1% Real	1% Real + synthetic	10% Real	10% Real + synthetic	20% Real	1% Real	1% Real + synthetic	10% Real	10% Real + synthetic	20% Real
CLAM-SB	0.725	**0.812**	0.886	**0.898**	0.91	0.603	**0.644**	0.649	**0.765**	0.787
DSMIL	0.609	**0.659**	0.838	**0.856**	0.905	0.517	**0.554**	0.563	**0.639**	0.669

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.

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.

Training Data	Test Acc
Real	78 %
Synthetic	70 %
Real + synthetic	83 %
SoTA [7]	**87 %**

(a)

Training Data	Val Acc
Real	93.8 %
Synthetic	90.19 %
Real + synthetic	**96.27 %**
SoTA [20]	96.26 %

(b)

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 $p(z \mid c)$ 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 $p(z \mid c)$. 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, 31] 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 [50]. 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 4×4 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) [33], we expect this drop in performance as we trained with machine-generated captions. Training and generated image-caption pairs are provided in supplementary.

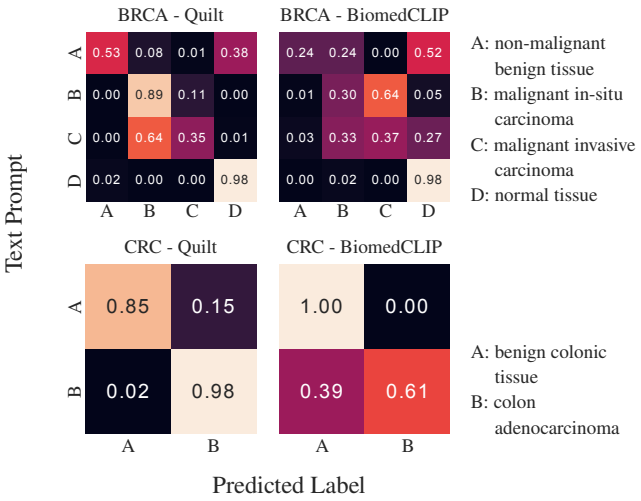

Figure 4. Confusion matrix of zero-shot classification for novel TCGA-CRC and TCGA-BRCA synthetic images.

7. Combining self-supervised embeddings with diffusion

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, 26] using features derived from our approach at $20\times$ and $5\times$ magnification ($\text{LDM}_{\text{HIPT/CTransPath}}$) 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.

Mag	Features	25% training CLAM-SB	25% training DSMIL	100% training CLAM-SB	100% training DSMIL
$20\times$	HIPT	0.788	0.784	0.861	0.839
	LDM_{HIPT}	**0.842**	**0.795**	**0.908**	**0.894**
$5\times$	CTransPath	0.900	0.896	0.919	0.910
	$\text{LDM}_{\text{CTransPath}}$	**0.913**	**0.905**	**0.923**	**0.936**

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

8. Conclusion

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.

Acknowledgements This research was partially supported by NCI awards 5U24CA215109, 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

[1] Marco Aversa, Gabriel Nobis, Miriam Hägele, Kai Standvoss, Mihaela Chirica, Roderick Murray-Smith, Ahmed Alaa, Lukas Ruff, Daniela Ivanova, Wojciech Samek, et al. Diffinfinite: Large mask-image synthesis via parallel random patch diffusion in histopathology. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 3

[2] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 3

[3] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, pages 1737–1752. PMLR, 2023. 4, 5

[4] JN Cancer Genome Atlas Research Network et al. The cancer genome atlas pan-cancer analysis project. *Nat. Genet*, 45 (10):1113–1120, 2013. 2, 4

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 1, 3, 4

[6] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16144–16155, 2022. 1, 3, 4, 8, 14

[7] Sai Saketh Chennamsetty, Mohammed Safwan, and Varghese Alex. Classification of breast cancer histology image using ensemble of pre-trained neural networks. In *Image Analysis and Recognition: 15th International Conference, ICIAR 2018, Póvoa de Varzim, Portugal, June 27–29, 2018, Proceedings 15*, pages 804–811. Springer, 2018. 7

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1, 2

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[10] Miguel Espinosa and Elliot J Crowley. Generate your own scotland: Satellite image generation conditioned on maps. *arXiv preprint arXiv:2308.16648*, 2023. 3

[11] Alexandre Filiot, Ridouane Ghermi, Antoine Olivier, Paul Jacob, Lucas Fidon, Alice Mac Kain, Charlie Saillard, and Jean-Baptiste Schiratti. Scaling self-supervised learning for histopathology with masked image modeling. *medRxiv*, pages 2023–07, 2023. 1, 3, 4, 7

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

[13] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 3

[14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5

[15] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 1, 2

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 14

[17] Vincent Tao Hu, David W Zhang, Yuki M Asano, Gertjan J Burghouts, and Cees GM Snoek. Self-guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18413–18422, 2023. 2

[18] Wisdom Oluchi Ikezogwo, Mehmet Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Stefan Chan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. Quilt-1m: One million image-text pairs for histopathology. *arXiv preprint arXiv:2306.11207*, 2023. 4, 7, 14, 17, 18

[19] Jakob Nikolas Kather, Niels Halama, and Alexander Marx. 100,000 histological images of human colorectal cancer and healthy tissue, 2018. 5, 7

[20] Anurodh Kumar, Amit Vishwakarma, and Varun Bajaj. Crccn-net: Automated framework for classification of colorectal tissue using histopathological images. *Biomedical Signal Processing and Control*, 79:104172, 2023. 7

[21] Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328, 2021. 6, 8

[22] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 4, 8, 14

[23] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022. 7

[24] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*, 2023. 1

[25] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pages 423–439. Springer, 2022. 14

[26] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient

and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570, 2021. 6, 8

[27] Puria Azadi Moghadam, Sanne Van Dalen, Karina C Martin, Jochen Lennerz, Stephen Yip, Hossein Farahani, and Ali Bashashati. A morphology focused diffusion probabilistic model for synthesis of histopathology images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2000–2009, 2023. 2

[28] Gustav Müller-Franzes, Jan Moritz Niehues, Firas Khader, Soroosh Tayebi Arasteh, Christoph Haarburger, Christiane Kuhl, Tianci Wang, Tianyu Han, Teresa Nolte, Sven Nebelung, et al. A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis. *Scientific Reports*, 13 (1):12098, 2023. 2

[29] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 1, 2

[30] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022. 1

[31] David Nukrai, Ron Mokady, and Amir Globerson. Text-only training for image captioning using noise-injected clip. *arXiv preprint arXiv:2211.00575*, 2022. 7

[32] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *CVPR*, 2022. 5

[33] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2, 8

[34] António Polónia, Catarina Eloy, and Paulo Aguiar. BACH Dataset : Grand Challenge on Breast Cancer Histology images, 2020. 5, 6

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4, 5, 14

[36] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 1

[37] Caleb Robinson, Le Hou, Kolya Malkin, Rachel Soobitsky, Jacob Czawlytko, Bistra Dilkina, and Nebojsa Jojic. Large scale high-resolution land cover mapping with multi-resolution data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12726–12735, 2019. 4

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2, 3, 5, 14

[39] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. 5

[40] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022. 2

[41] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 1

[42] Ahmad Sebaq and Mohamed ElHelw. Rsdiff: Remote sensing image generation from text using diffusion model. *arXiv preprint arXiv:2309.02455*, 2023. 3

[43] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020. 2, 14

[44] USGS. National agriculture imagery program (NAIP), 2023. https://www.usgs.gov/centers/eros/science/usgs-eros-archive-aerial-photography-national-agriculture-imagery-program-naip. 2, 4

[45] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Junzhou Huang, Wei Yang, and Xiao Han. Transpath: Transformer-based self-supervised learning for histopathological image classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*, pages 186–195. Springer, 2021. 4, 6

[46] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical image analysis*, 81:102559, 2022. 1, 3

[47] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16133–16142, 2023. 6

[48] Xuan Xu, Saarthak Kapse, Rajarsi Gupta, and Prateek Prasanna. Vit-dae: Transformer-driven diffusion autoencoder for histopathology image analysis. *arXiv preprint arXiv:2304.01053*, 2023. 2

[49] Srikar Yellapragada, Alexandros Graikos, Prateek Prasanna, Tahsin Kurc, Joel Saltz, and Dimitris Samaras. PathLDM: Text conditioned latent diffusion model for histopathology. *arXiv preprint arXiv:2309.00748*, 2023. 2, 4, 5, 6

[50] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, et al. Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv preprint arXiv:2303.00915*, 2023. 7

Learned representation-guided diffusion models for large-image generation

Supplementary Material

9. Annotation costs

We asked a pathologist to annotate patches from TCGA-BRCA to estimate the cost of detailed per-patch annotation for the entirety of the dataset. We presented the $20\times$ magnification patches of Fig. 5 and requested them to *"write a brief description for each of the following patches"*. An expert pathologist required approximately 5-10 seconds to identify features and describe the patches. Therefore, for the entire 15M patches of TCGA-BRCA, it would take ≈ 40000 hours to provide full per-patch annotations. This training dataset is small compared to the volume of data used in large studies, e.g. 10k whole slide images or approximately $10\times$ the number of TCGA-BRCA data. Employing expert pathologists to annotate these vast amounts of data is prohibitively expensive and, therefore, practically infeasible at the scale at which we want to apply these models.

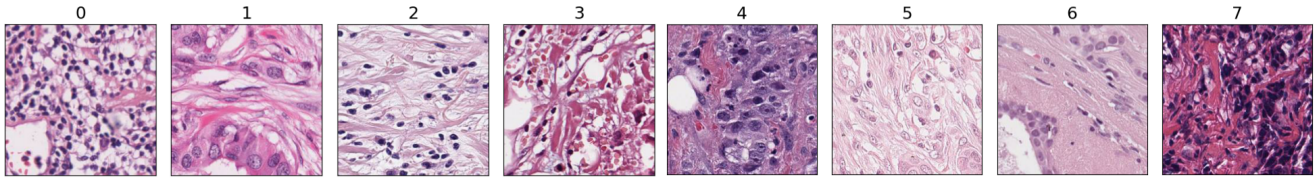

Figure 5. Examples of patches annotated by an expert pathologist. For each image, the pathologist required 5-10s to provide a brief, detailed description of the features visible. Annotating the entirety of TCGA in this manner is a colossal task.

10. Out-of-distribution augmentation examples

In Fig. 6 and Fig. 7 we show out-of-distribution examples of generated images from the NCT-CRC and BACH datasets, along with the reference image from which the SSL embeddings were extracted. For NCT-CRC, it is evident that the synthetic patches follow the semantics and appearance of the real patches used. Regarding BACH, we find the appearance to be slightly different between the real large images and our synthetic large images, but we see that the semantic contents are mostly left unchanged. This is also validated by our augmentation experiments in the main text, where we improve the classification accuracy with synthetic BACH data. In both cases, our SSL-conditioned diffusion models exhibit impressive generalization capabilities by only modifying the conditioning provided to them. Given that generalization is an essential property for building foundation models, we believe that our work is an important step towards this direction for large image domains such as digital histopathology and remote sensing.

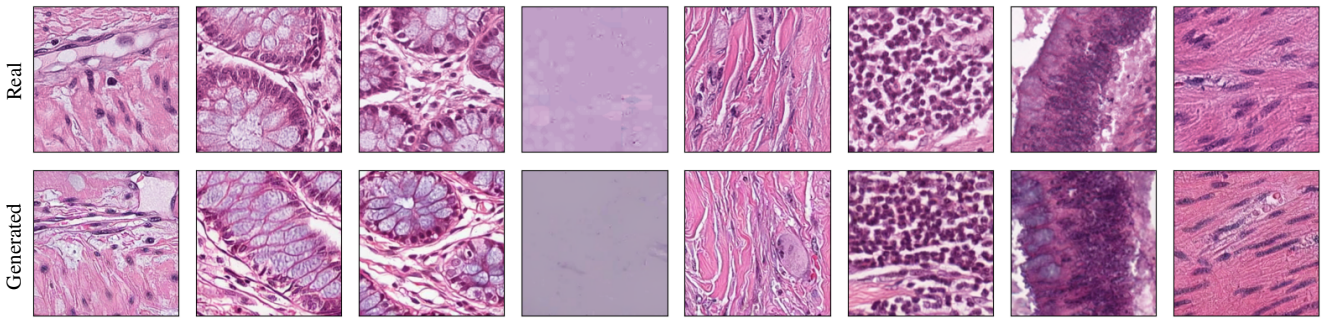

Figure 6. Synthetic images from NCT-CRC. For each generated image we extract the SSL embedding from a real reference image, taken from NCT-CRC-HE-100K, and generate a patch using the TCGA-CRC model. The synthesized patches are similar to the reference in both appearance and semantics. The TCGA-CRC model was never trained on data from the NCT-CRC-HE-100K dataset.

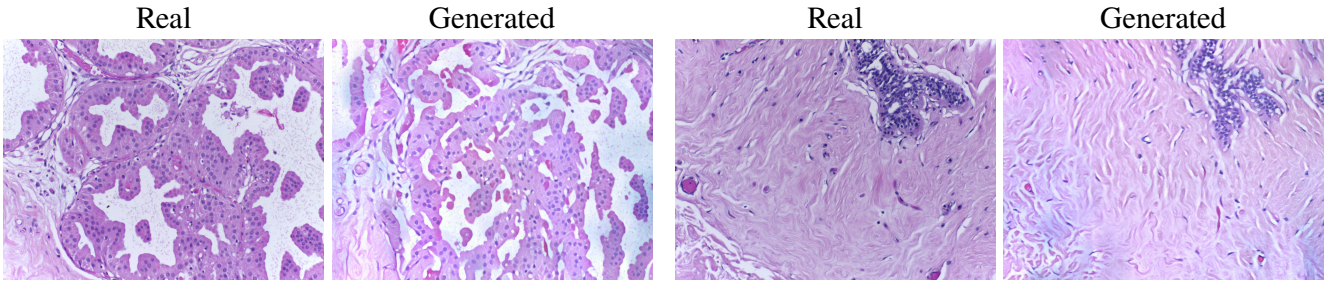

Figure 7. Examples of generated images from BACH. For each generated image we extract the SSL embeddings from a reference image, taken from the BACH dataset. We generate the large image using the TCGA-BRCA model. Although the appearance between the reference and generated images is slightly different, the large images maintain the global semantics. The TCGA-BRCA model was never trained on data from the BACH dataset.

11. NCT-CRC augmentation additional results

We expand the results of Table 3b in the main text by evaluating the classification accuracy on the CRC-VAL-HE-7K test set with more synthetic data. As shown in Table 5, expanding the dataset with more than $2\times$ synthetic data does not improve the performance further. Adding more synthetic data ends up hurting the classifier, which we attribute to the dilution of the real data with the imperfect, synthetic variations that we generate with our diffusion model. Even so, the final classification accuracy with $5\times$ synthetic data is still higher than the baseline that only uses real images.

Training Data	Val Acc
Real	93.8%
Real + $1\times$ Synthetic	96.27%
Real + $2\times$ Synthetic	**96.55%**
Real + $3\times$ Synthetic	95.52%
Real + $5\times$ Synthetic	95.59%

Table 5. Classification accuracy on the CRC-VAL-HE-7K test set for different quantities of synthetic data. Expanding the training data with more than $2\times$ synthetic samples does not improve the classification accuracy further.

12. Memory requirements

We propose using an LDM trained on 256×256 pixel patches to generate large images of size 1024×1024. In Table 6, we compare the requirements of training an LDM directly on 1024×1024 pixel large images, instead of our patch-based approach. We use the same $4\times$ downsampling factor for the first stage VAE and employ a single RTX 6000 GPU for benchmarking purposes. Training a diffusion model on 1024×1024 resolution TCGA-BRCA images requires an order of magnitude more time than our patch-based approach for the same number of iterations. However, with a reduced batch size, we also empirically know that it would require more training iterations for the model to converge. We argue that since our approach can be used to generate large images without significant loss in quality, our patch-based model is the more efficient solution.

Training Method	Maximum batch size	Training time per epoch
Ours (256×256 patches)	100	45 hr
LDM on 1024×1024 images	4	300 hr

Table 6. Training a diffusion model on large images is computationally expensive and takes an order of magnitude more time.

13. Using different SSL encoders

We extend the TCGA-BRCA $20\times$ model of Table 1 with additional patch-level FID values, obtained by using different embeddings as conditioning (Table 7). The pathology-specific HIPT performs best, suggesting that the domain expressivity

Conditioning	Patch FID
None	25.62
ImageNet ViT-B/16	13.29
CLIP [35]	16.07
HIPT [6]	**6.98**

Table 7. FIDs when using different representations as conditions.

Stride	Time/ Image	Crop FID	CLIP FID
4	15m	**12.66**	**7.31**
8	4m	14.69	7.37
16*	1m	15.51	7.43
32	20s	15.60	8.09

Table 8. Large image generation parameters ablation. By * we denote the stride used in the main text experiments.

of the embedding used as conditioning affects image generation quality. We conjecture that worse patch quality also hurts large image metrics.

14. Large image generation details

To generate large images we use DDIM [43] with 50 inference steps and a classifier-free guidance weight of 3.0. The SSL conditioning (384 or 768 dimensional vector depending on the SSL model) is first normalized with the L_2 norm and then projected to a 512-dim vector using a linear layer. The null token for the classifier-free guidance is represented by replacing the SSL embedding with a vector of all 0s. The conditioning is applied to the U-Net model using cross-attention, similar to other LDM conditioning mechanisms [38].

The LDM is applied to patches in the large image with a stride of 16. Using a larger stride leads to tiling artifacts, whereas a smaller stride increases the computational cost without much difference in the synthesized image quality. In Table 8 we provide an ablation study of the large image generation parameters. We synthesize 1024×1024 px images from TCGA-BRCA with different strides, using 50 steps of diffusion, on an NVIDIA RTX 6000, showing that larger strides require fewer forward passes (less time) but produce worse results.

For each location i, j at which we want to apply the diffusion model, we interpolate the 4-nearest embeddings to get the conditioning $\lambda_{i,j}$. We found that spherical linear interpolation (slerp), weighted by the distance of i, j to the centers of its four neighbors, worked best for interpolating the high-dimensional, normalized SSL embeddings.

When averaging the diffusion updates we first applied a Gaussian kernel to downweight the pixels at the edges of the patch. This helps with unwanted tiling artifacts as we 'trust' the diffusion updates in the center more than the edges of a patch. Likewise, when decoding the large image latents into images, we used a stride of 16 with the Gaussian kernel weighting, to eliminate tiling artifacts in the decoded images.

For the text-to-large image experiments, we trained an auxiliary diffusion model to sample a 4×4 grid of embeddings given the text conditioning. We used a small convolutional network with residual layers to implement the diffusion model. The timestep conditioning was concatenated to the input grid of embeddings. The network directly predicted the final embeddings from the conditioning and current noisy embedding grid, instead of predicting the noise added. For TCGA-BRCA and TCGA-CRC, the text conditioning is a single 512-dim Quilt embedding vector. For NAIP, we used a frozen CLIP [35] text encoder to extract features from the text captions and used them as conditioning. For the diffusion process, we used 1000 steps with a linear schedule, as in [16]. Additionally, when sampling embeddings from text for TCGA-BRCA and TCGA-CRC we used negative prompting [25] to further separate the different types of images during generation.

15. Text-to-large image generation examples

In Fig. 10 we present generated images from the TCGA-BRCA model and the text prompts used in generating them. We borrow the text prompts from the zero-shot classification experiments of [18]. As discussed for the confusion matrix (Fig. 4), the vision-language model's capabilities limit the quality of our results. The model seems to be able to only differentiate between *non-malignant* / *normal* and *malignant*, which is expected since the zero-shot classification accuracy of Quilt on breast cancer images is around 40%. In contrast, for the CRC data where accuracy is around 90%, our text-to-large image generation performs better. In Fig. 11 we present such synthetic samples from TCGA-CRC.

To train the satellite text-to-large image auxiliary diffusion model we generated a synthetic set of image-caption pairs using BLIP [22]. For training, we created a set of 30k large images (1024×1024 pixels) with 4 captions for each, whereas for the test set, we used a single caption for evaluation. In Fig. 12 we present images from the training and test sets as well as generated samples along with their text prompts. We see that although the training captions are far from perfect, we are able to generate test set images consistent with the prompts used. Even though our training set is tiny, we see interesting

generalization capabilities when using 'unusual' prompts, such as *"a satellite image of a forest with smoke"*, where the model tries to add clouds to mimic the "smoke" seen from a satellite image. This generalization can be attributed to both the expressivity of the SSL embeddings used in synthesizing the images and the usage of a pre-trained CLIP text encoder to interpret the captions.

16. Pathologist evaluation

We designed a simple user interface where we presented large TCGA-CRC images generated from text prompts and asked an expert pathologist to evaluate them (Fig. 8). The model generated an image using one of two text prompts: *"benign colonic tissue"* or *"colon adenocarcinoma"*. We asked a pathologist to evaluate by categorizing the images as *benign* / *adenocarcinoma* / *undecided* as well as assigning a *realistic* / *unrealistic* label. For a total of 100 images, the final agreement between text prompts and pathologist labels was 89.9%, with 61% of the images marked as realistic. This clearly illustrates the applicability of our proposed method; an auxiliary diffusion model that generates the SSL conditioning from any related modality can be chained with our patch-based diffusion to synthesize coherent large images.

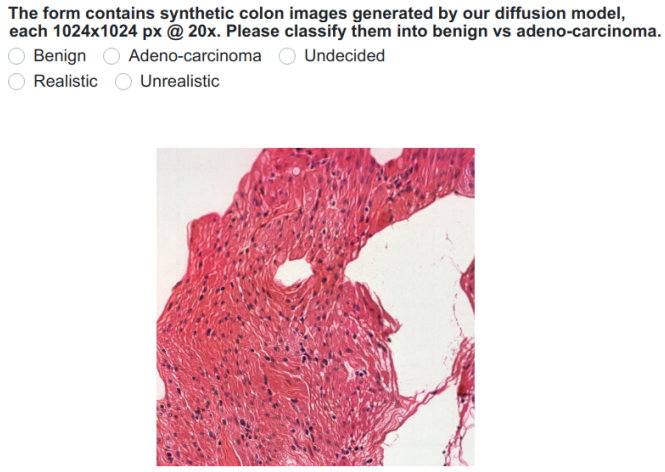

Figure 8. Pathologist evaluation UI. We presented synthetic images to an expert pathologist and asked them to evaluate them. The results showed 89.9% agreement between the text prompts used to generate the images and the pathologist's assessment.

17. Embedding resolution

In Fig. 9 we show synthetic large images, using different embedding granularities from a reference image. When utilizing the full embedding resolution, we use the entire 4×4 embedding grid to generate a variation of the original image by interpolating to get conditioning at each i, j location. At half resolution, we average the embeddings and use a 2×2 grid, leading to more repeated textures in the final image. When using a single embedding (patch indicated with a green box) the generated image is equivalent to infinitely tiling the textures from the reference patch.

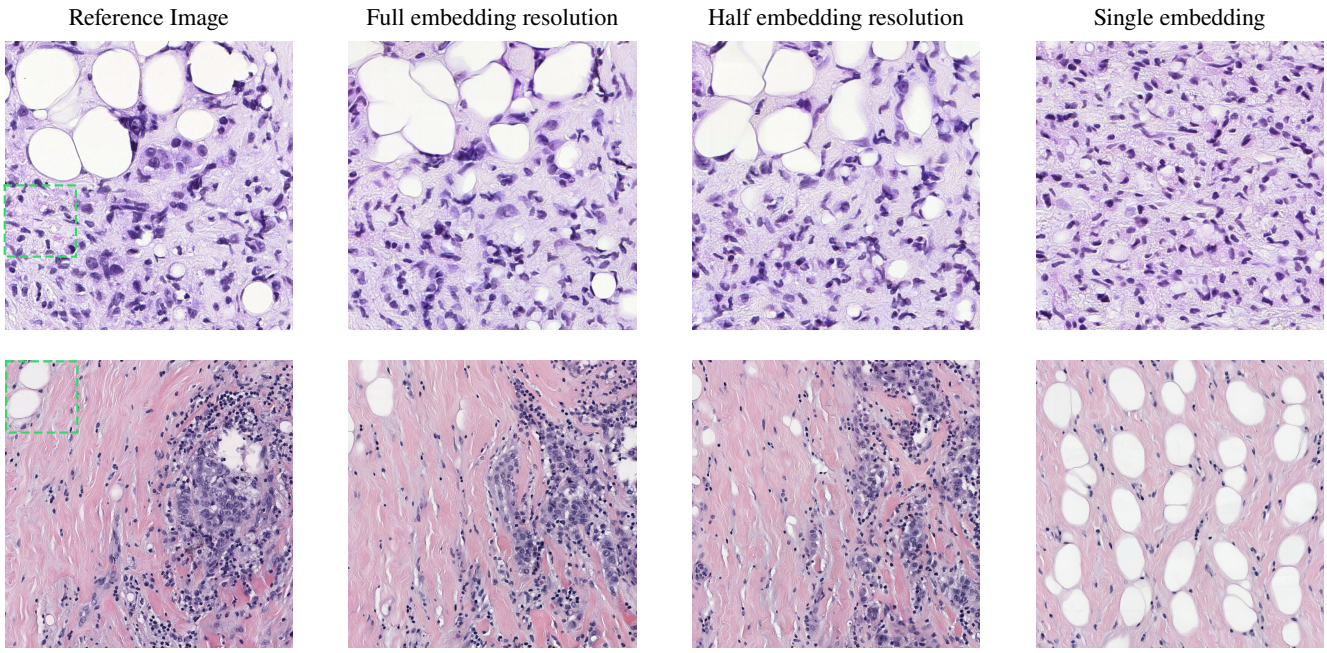

Figure 9. Using coarser conditioning results in repeated textures in the generated large image. When using a single embedding the result is equivalent to an infinitely-tiled patch. Images are at 1024×1024 pixels resolution.

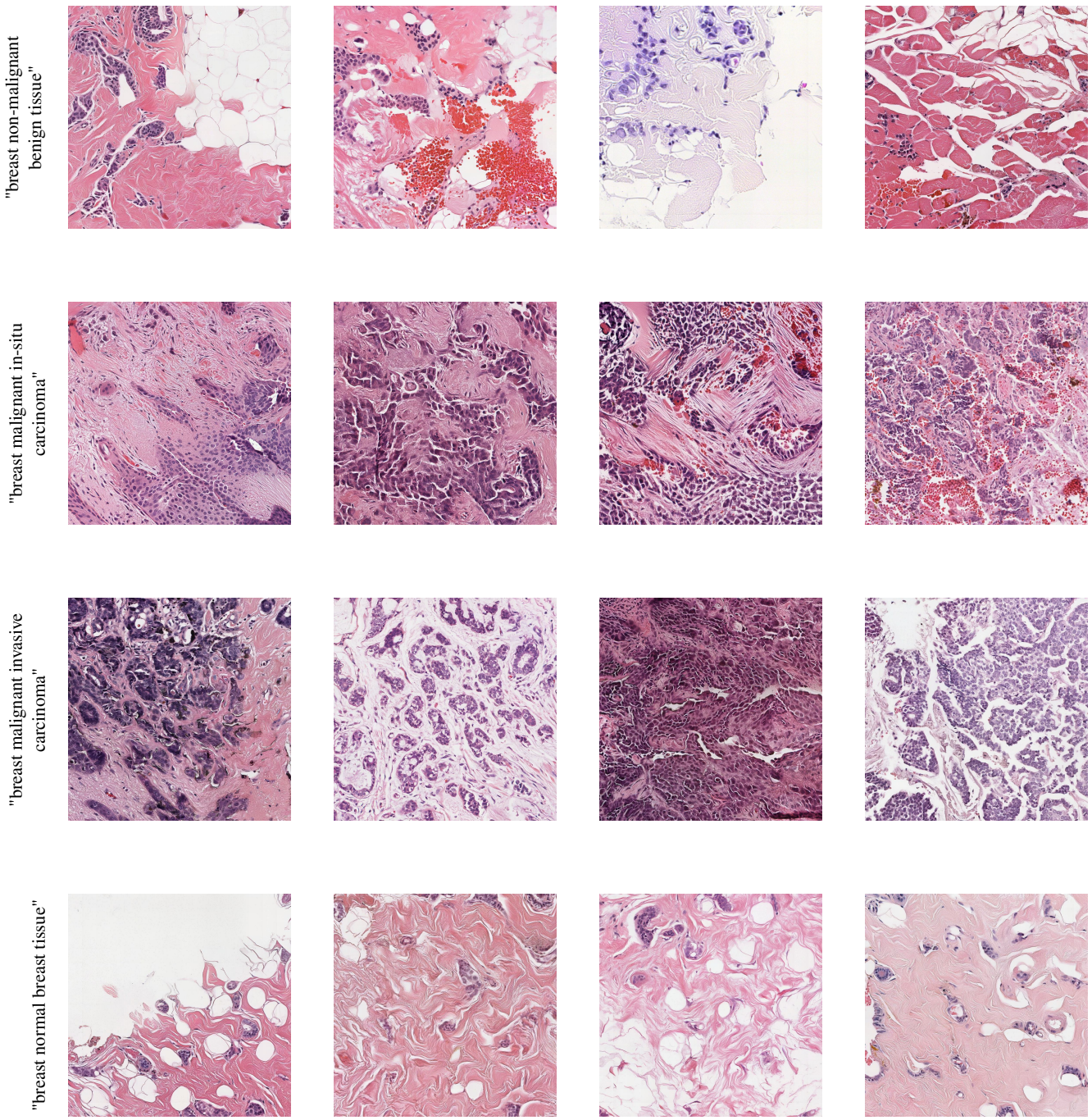

Figure 10. Generated samples from TCGA-BRCA along with the text prompt used. We use the zero-shot classification prompts from Quilt [18] to generate the embeddings. Images are at 1024×1024 pixels resolution.

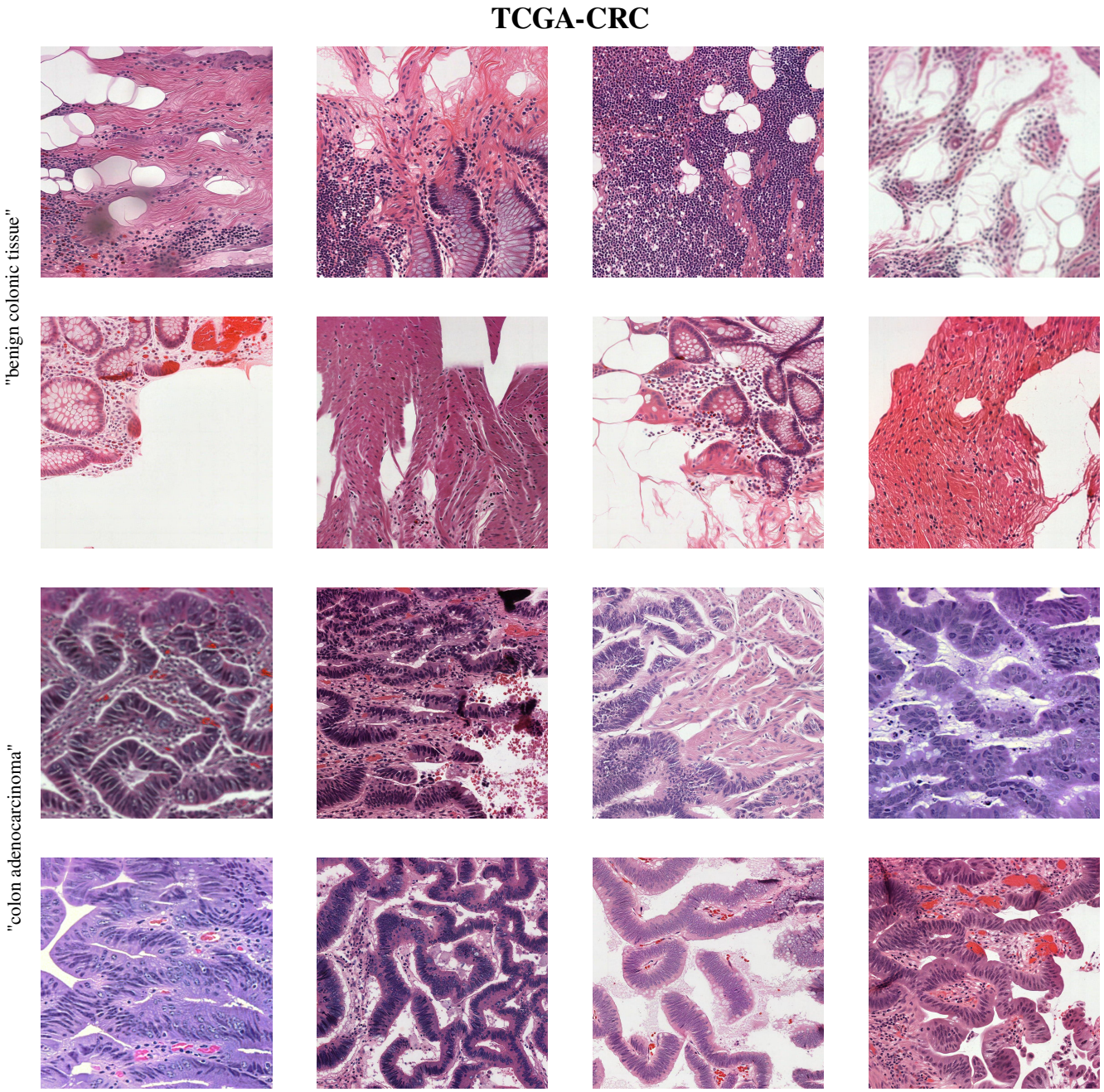

Figure 11. Generated samples from TCGA-CRC along with the text prompt. We use the zero-shot classification prompts from Quilt [18] to generate the embeddings. Images are at 1024×1024 pixels resolution.

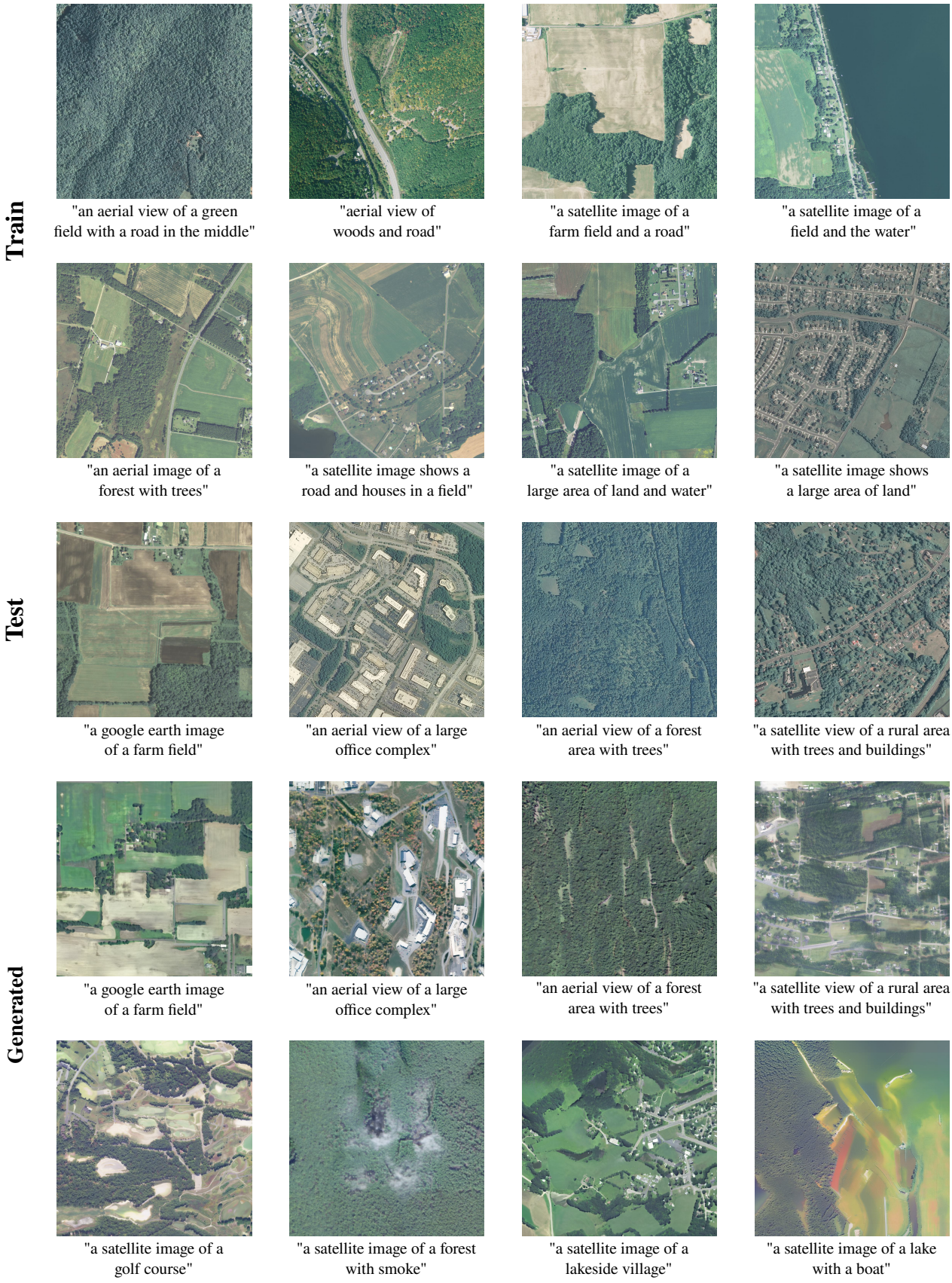

Figure 12. Examples of training, test and generated text-to-large satellite images. Images are at 1024×1024 pixels resolution.