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. 6 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.

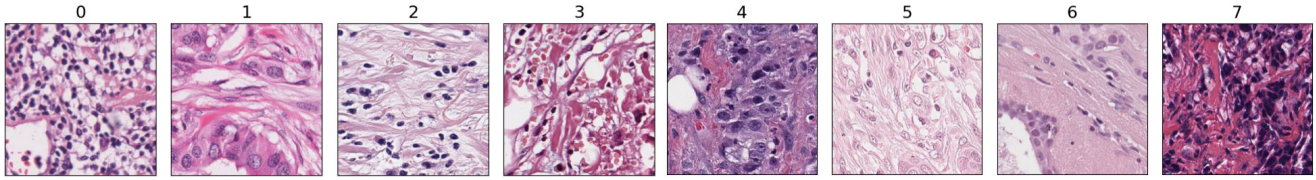

Figure 6. Examples of patches annotated by an expert pathologist. For each image, the pathologist required 5-10s to provide a brief, detailed description of the features visible. Annotating the entirety of TCGA in this manner is a colossal task.

10. Out-of-distribution augmentation examples

In Fig. 7 and Fig. 8 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.

11. NCT-CRC augmentation additional results

We expand the results of Table 3(b) 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.

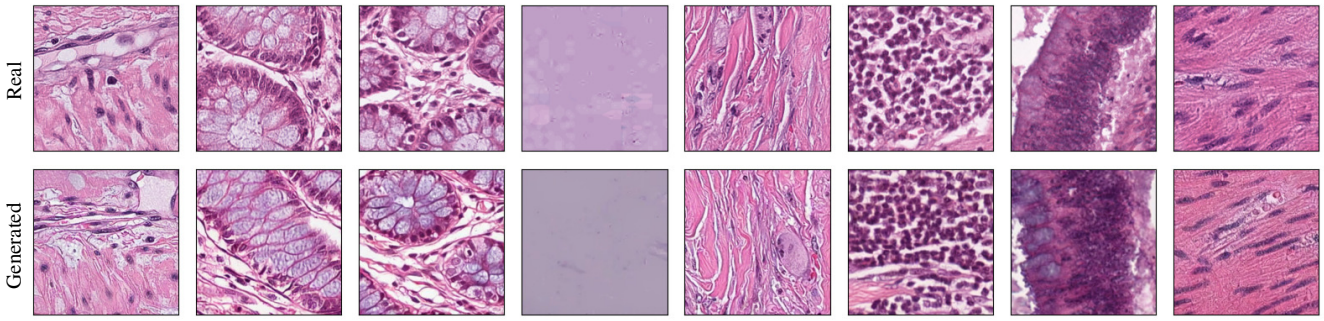

Figure 7. Synthetic images from NCT-CRC. For each generated image we extract the SSL embedding from a real reference image, taken from NCT-CRC-HE-100K, and generate a patch using the TCGA-CRC model. The synthesized patches are similar to the reference in both appearance and semantics. The TCGA-CRC model was never trained on data from the NCT-CRC-HE-100K dataset.

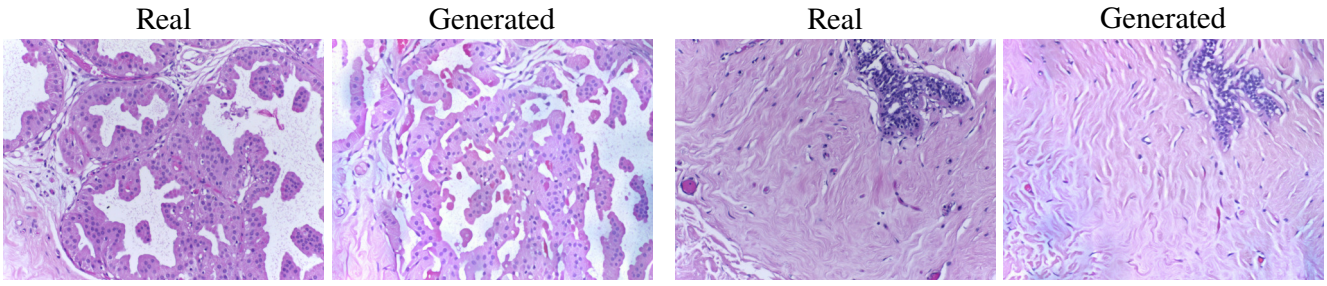

Figure 8. Examples of generated images from BACH. For each generated image we extract the SSL embeddings from a reference image, taken from the BACH dataset. We generate the large image using the TCGA-BRCA model. Although the appearance between the reference and generated images is slightly different, the large images maintain the global semantics. The TCGA-BRCA model was never trained on data from the BACH dataset.

Training Data	Val Acc
Real	93.8%
Real + 1× Synthetic	96.27%
Real + 2× Synthetic	**96.55%**
Real + 3× Synthetic	95.52%
Real + 5× 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× 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.

Conditioning	Patch FID
None	25.62
ImageNet ViT-B/16	13.29
CLIP [6]	16.07
HIPT [1]	**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.

13. Using different SSL encoders

We extend the TCGA-BRCA 20× model of Table 1 (main paper) 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 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 [8] 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 [7].

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 [6] 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 [2]. Additionally, when sampling embeddings from text for TCGA-BRCA and TCGA-CRC we used negative prompting [5] to further separate the different types of images during generation.

15. Text-to-large image generation examples

In Fig. 11 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 [3]. As discussed for the confusion matrix (Fig. 4 in the main text), 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. 12 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 [4]. 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. 13 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. 9). 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.

17. Embedding resolution

In Fig. 10 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.

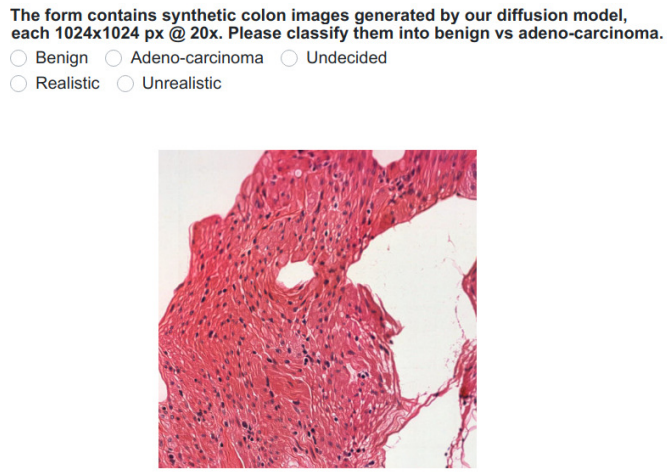

Figure 9. Pathologist evaluation UI. We presented synthetic images to an expert pathologist and asked them to evaluate them. The results showed 89.9% agreement between the text prompts used to generate the images and the pathologist's assessment.

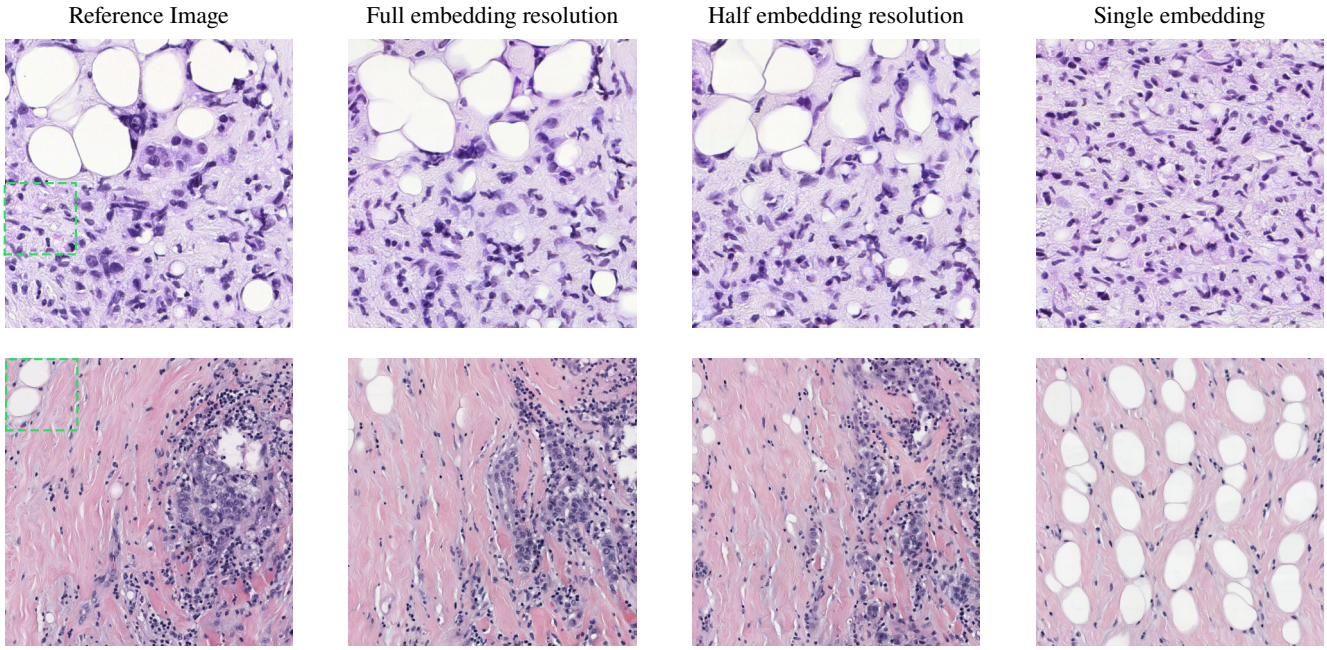

Figure 10. Using coarser conditioning results in repeated textures in the generated large image. When using a single embedding the result is equivalent to an infinitely-tiled patch. Images are at 1024×1024 pixels resolution.

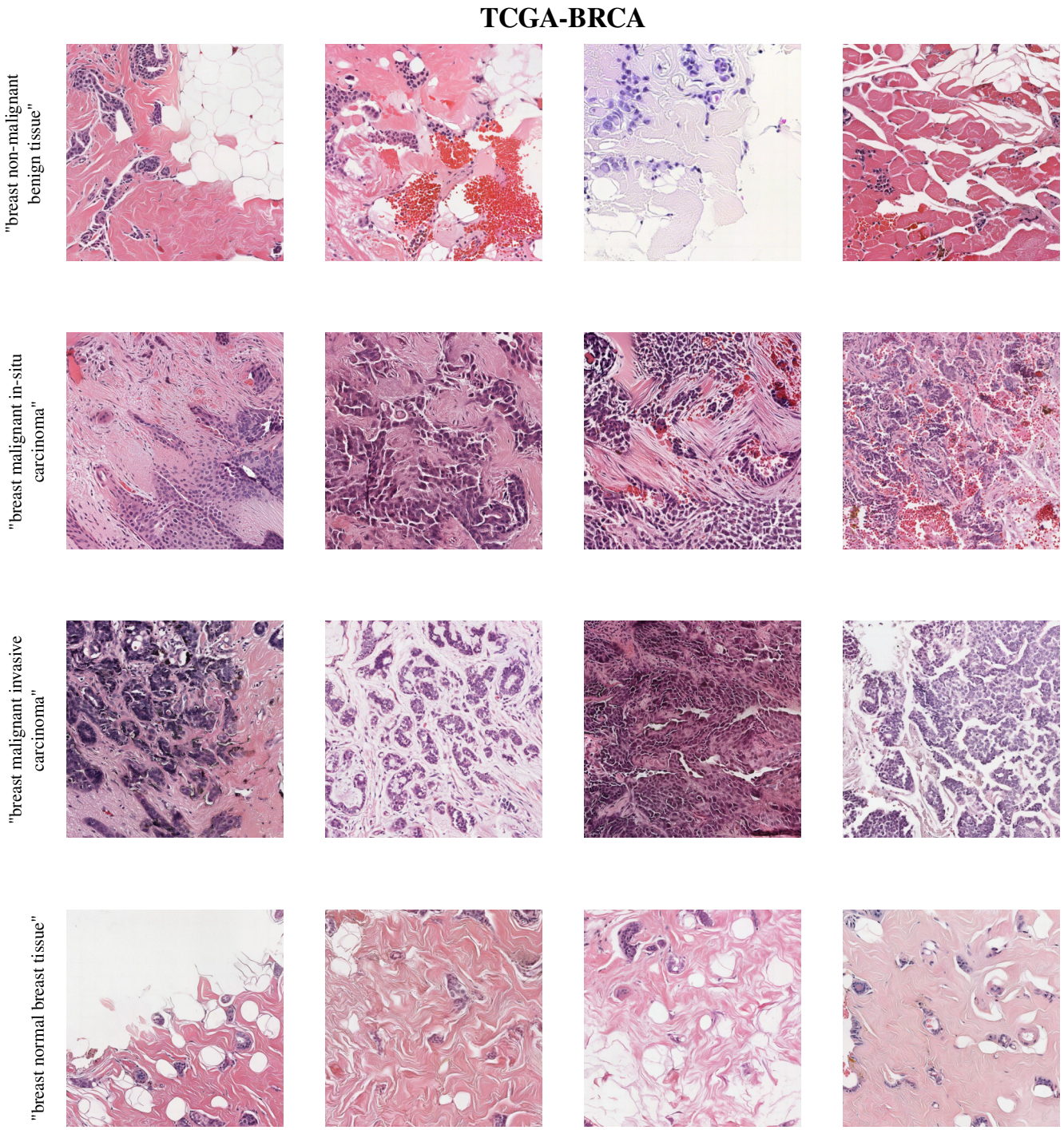

Figure 11. Generated samples from TCGA-BRCA along with the text prompt used. We use the zero-shot classification prompts from Quilt [3] to generate the embeddings. Images are at 1024×1024 pixels resolution.

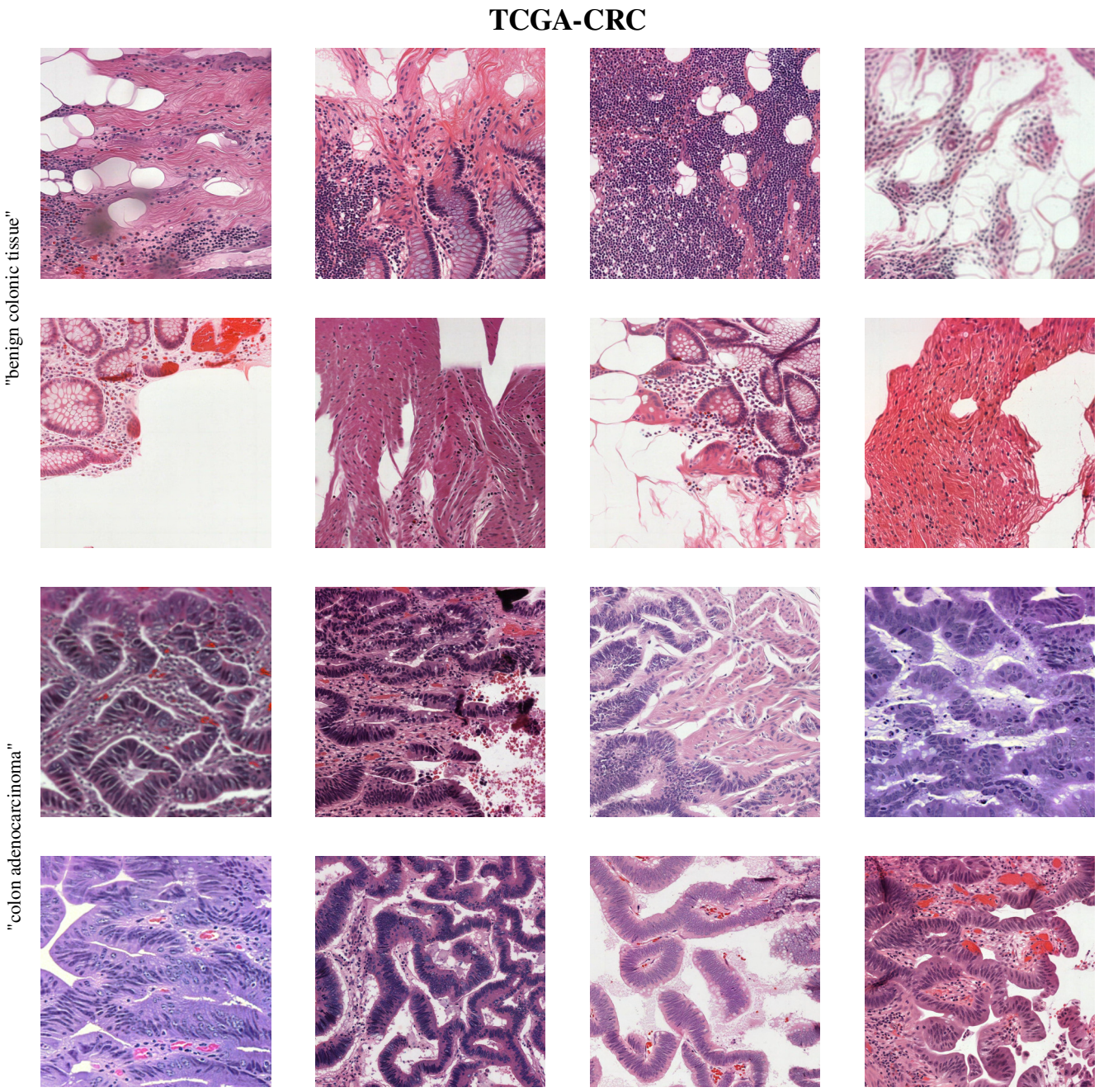

Figure 12. Generated samples from TCGA-CRC along with the text prompt. We use the zero-shot classification prompts from Quilt [3] to generate the embeddings. Images are at 1024×1024 pixels resolution.

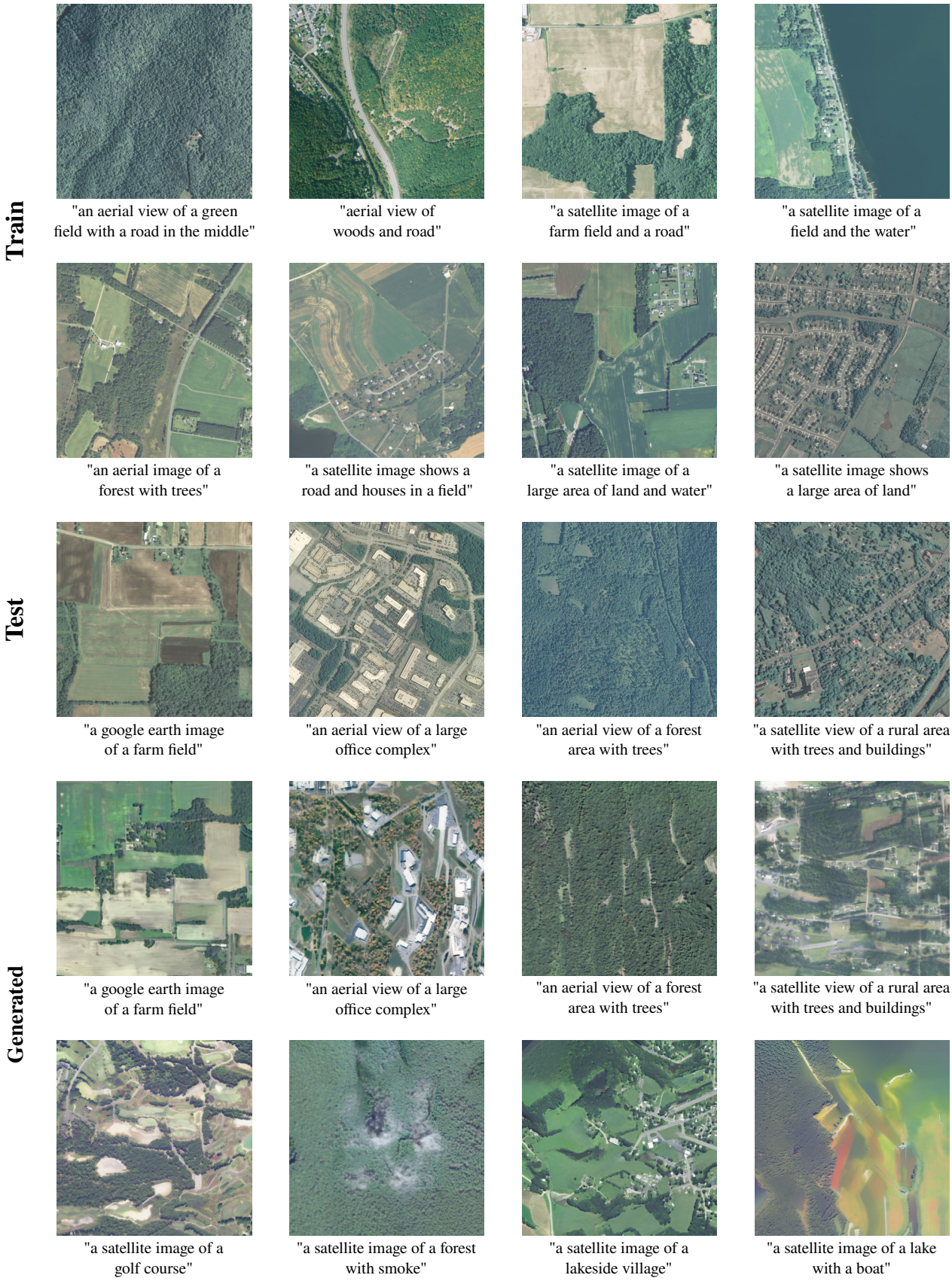

Figure 13. Examples of training, test and generated text-to-large satellite images. Images are at 1024×1024 pixels resolution.

References

[1] 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. 3

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 4

[3] 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, 6, 7

[4] 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

[5] 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. 4

[6] 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. 3

[7] 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. 3

[8] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020. 3