PathDiff: A Diffusion Framework for Histopathology Image Synthesis Unifying Unpaired Text and Mask Conditions

Technical Report

Student: Abdul Wasi
Advisor: David Doermann
Department of Computer Science and Engineering
University at Buffalo, Buffalo, NY, USA

Abstract

Diffusion-based generative models have shown considerable promise in synthesizing histopathology images to address data scarcity driven by privacy constraints. Text and masks are complementary modalities that provide valuable insights beyond the images, serving as effective conditions for generative models. Text-based diagnostic reports offer high-level semantic descriptions but may lack structural specificity, while masks provide pixel-level spatial information essential for representing distinct regions or structures. Integrating both for image generation is highly beneficial, as it allows for precise control over high-level semantics and fine-grained spatial details while maximizing available data—a particularly valuable strategy given the data scarcity. Unfortunately, no publicly available histopathology datasets contain paired diagnostic reports and masks for the same images, presenting a significant challenge for scalable data generation. To address this problem, we propose PathDiff, a diffusion framework that jointly learns from complementary, unpaired mask-text data. By integrating both modalities into a single conditioning space, our model maximizes the utility of limited data and enables generation with structural and contextual guidance during inference. Extensive experiments show that our framework outperforms existing methods in image fidelity, image-text alignment, image-mask alignment, and downstream segmentation tasks. Our code and models will be open-sourced.

Student's Contribution

This technical report presents my contribution to the collaborative research paper "PathDiff: A Diffusion Framework for Histopathology Image Synthesis Unifying Unpaired Text and Mask Conditions".

My contributions to this work include participating in the design and implementation of the PathDiff framework, designing the main pipeline, conducting and analyzing experiments on the mask-to-image and text-to-image generation tasks, evaluating performance across multiple histopathology datasets (PanNuke, CoNIC, MoNuSAC, and PathCap), and contributing to the preparation of the manuscript.

Contributions of Co-Authors

This work is the product of a collaborative team effort. The following co-authors contributed to the research and development of PathDiff:

- **Mahesh Bhosale** (University at Buffalo): Research leadership, and manuscript preparation.
- **Yuanhao Zhai** (University at Buffalo): Experimental design and evaluation methodology.
- **Yunjie Tian** (University at Buffalo): Data processing pipeline and experimental analysis.
- **Nan Xi** (University at Buffalo): Model implementation and ablation studies.
- **Samuel Border** (University of Florida): Pathology domain expertise and domain expert survey coordination.
- **Pinaki Sarder** (University of Florida): Histopathology supervision and expert evaluation oversight.
- **Xuan Gong** (Harvard Medical School): Research conceptualization, technical direction, and manuscript guidance.
- **David Doermann** (University at Buffalo): Faculty advising, research oversight, and manuscript review.

1. Introduction

The recent advancements in computational pathology, driven by deep learning, are transforming the field of histopathology by addressing critical challenges in tasks such as nuclei classification and segmentation [39], survival prediction [41], multi-instance learning [11], and transfer learning [39]. Despite these successes, a major obstacle persists: the substan-

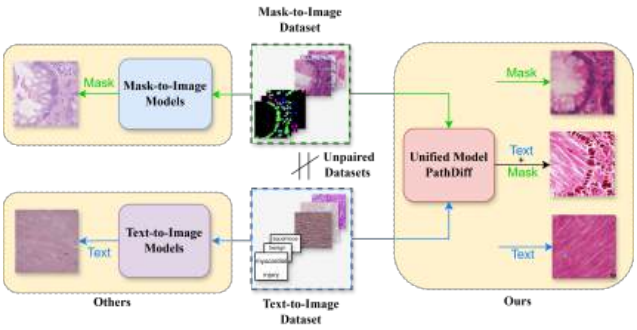


Figure 1. **PathDiff is trained on unpaired datasets, integrating two conditional modalities —*Text* and *Mask*— to enable versatile image generation.** Unlike unimodal conditional models, PathDiff can generate images conditioned on *Text*, *Mask*, or *both*, allowing greater control and adaptability in image synthesis.

tial volume of annotated data necessary to effectively train deep learning models. Moreover, the high costs and domain expertise required to annotate such data further aggravate this problem. Pathologists must start with low magnification to assess tissue architecture and cellular arrangement, then shift to higher magnification to evaluate finer details such as cell morphology, nucleoli appearance, and chromatin density. Annotating these intricate features is both time-consuming and labor-intensive. For instance, fully annotating the 1k whole slide images from TCGA dataset[1] [9] would require approximately 40k pathologist hours [16]. Overcoming these limitations is essential to fully unlocking the potential of deep learning in histopathology [42]

Generative models in histology [8, 29] have thus emerged as a valuable tool to supplement existing datasets, extending beyond traditional data augmentation [10, 27]. More recently, owing to the superior generation quality of diffusion models [6, 19], they are widely used for histology image synthesis, either conditioned on diagnostic text reports [16, 45] or spatial labels like cell nuclei [31, 32] or regions of interest [1]. However, these existing approaches rely solely on a single modality for conditioning, which limits both the quality of control and the amount of data that can be utilized.

We recognize the importance of considering text and masks, as they provide complementary information. The text offers contextual knowledge, which varies across cancer types, grades, disease stages, or tissue abnormalities, enabling high-level semantic control over the generation process. In contrast, spatial masks capture local structural details, such as cell shapes and types, providing critical spatial information. Intuitively, simultaneous control of both contextual and spatial information enhances generation. However, no publicly available datasets include open-world text and spatial mask annotations alongside images.

We propose PathDiff, a diffusion framework that unifies text and mask conditions to address the abovementioned challenges. As shown in Fig. 2, it learns jointly from two independent datasets: one containing image-text pairs and the other containing image-mask pairs. By integrating unpaired text and mask into a single latent space, our model maximizes the use of complementary data. This enables more efficient sampling to generate images during inference, conditioned on text and mask. Our main contributions can be summarized as follows:

- To address the data scarcity issue in histology image analysis, we propose a novel diffusion-based framework that unifies unpaired text and masks conditions within a single latent space for complementary knowledge exploration.
- Our pipeline enables joint learning from independent text-image and mask-image datasets, enabling image generation conditioned on both during inference.
- Empirical evaluations demonstrate that our proposed model outperforms existing approaches across various evaluations, including image fidelity, image-text/mask alignment, and downstream classification and segmentation tasks.

2. Related Work

Generative adversarial networks [13] have become popular for medical image synthesis [3, 4, 35]; however, they frequently introduce artifacts, particularly in histopathology images [30]. In contrast, diffusion-based methods [19] have demonstrated enhanced image quality in both natural and medical images [30]. Building on this, Classifier Guidance [7] was introduced to condition image generation on specific classes, followed by Classifier-Free Guidance (CFG) [18], which eliminates the need for an auxiliary classifier. Latent Diffusion Models (LDMs) [37] further enhance the computational efficiency, while ControlNet [47] utilizes CFG to introduce multiple spatial controls in Text-to-Image LDMs [37]. Recent approaches for conditional histopathology image generation emphasize text- and mask-based conditioning, which we briefly review in the following sections.

2.1. Text to Histopathology Image Synthesis

The application of diffusion models in text-conditioned histopathology image generation remains limited. Summarized reports from the Large Text-Image histopathology dataset TCGA-BRCA [9] were used for text-conditioned image generation in [45] in conjunction with text annotated Tumor-Infiltrating Lymphocyte (TIL) and Tumor probabilities from off the shelf classifiers. Genome sequencing data from the TCGA-BRCA dataset was used in [29], where heavier weights were assigned to earlier steps and lower weights to later steps in the diffusion process to focus on morphological features. Authors in [16] train diffusion model conditioned on Self Supervised Learnt (SSL) embeddings. An auxiliary diffusion model is trained on SSL embedding

[1] https://www.cancer.gov/ccg/research/genome-sequencing/tcga

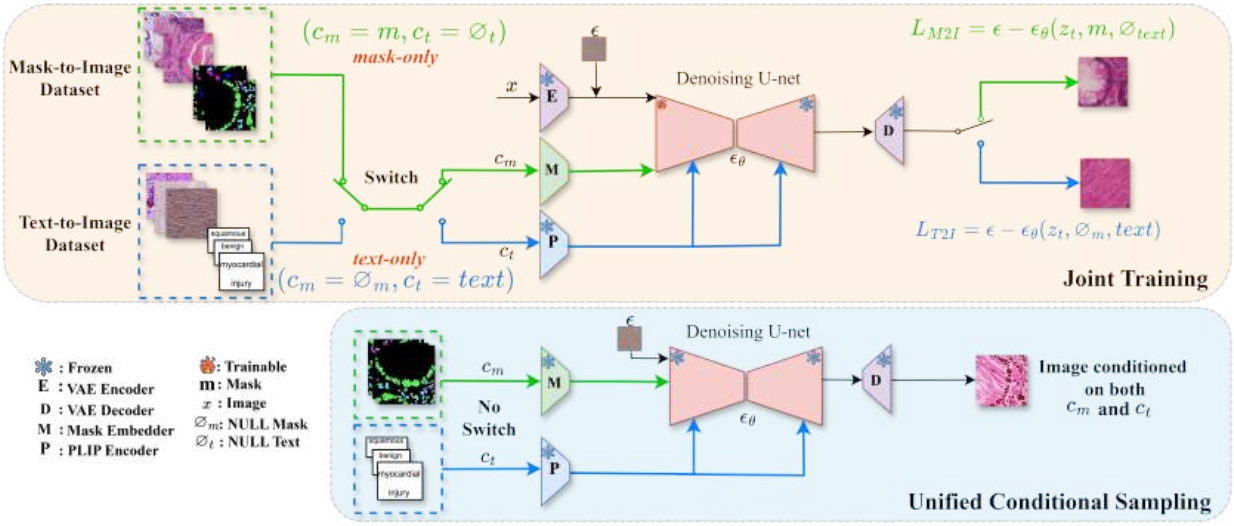


Figure 2. (a) **PathDiff Training Framework:** A training sample is drawn from either the Mask-to-Image (M2I) or the Text-to-Image (T2I) dataset, determined by the probability p_{split}, which controls the connection of the Switch. The missing condition in each case is set to $\varnothing$. When the Switch selects a sample from the M2I dataset, the Mask-to-Image loss L_{M2I} is applied; when it selects from the T2I dataset, the Text-to-Image loss L_{T2I} is used. This approach enables joint training of a single diffusion model on both datasets. (b) **Image Generation:** During generation, both conditions, c_m (mask) and c_t (text), are applied to produce samples that unify both conditions.

paired with Quilt [24] image embeddings. At the time of inference, text embeddings are used as proxies for image embedding allowing for text-to-image synthesis. However, note that none of these methods use mask conditions in conjunction with text.

2.2. Mask to Histopathology Image Synthesis

The authors of [1] propose a hierarchical diffusion model to generate large whole-slide images (WSIs) conditioned on synthesized regions of interest. But, this approach does not incorporate fine-grained, cell-level masks. In [32], a text-driven approach is used first to generate cell masks, which along with their distance maps are then used to condition histopathology image synthesis. However, this method does not allow for fine-grained control over the spatial placement of the masks. In [46], the diffusion model is trained to synthesize nuclei structures as pixel-level semantic and distance-transform maps, which are then post-processed into instance maps. This is followed by a conditional diffusion model to generate histopathology images. Similarly, [28] introduces a cell-point map to synthesize cell-type masks and images jointly. To tackle class imbalance, [31] employs a Semantic-Diffusion-Model (SDM) [44] for data synthesis, effectively balancing class variance in nuclei datasets.

In contrast, PathDiff uniquely generates histopathology images by unifying mask and text conditions from unpaired datasets, which is beyond the scope of the works discussed above.

3. Method

3.1. Background

Diffusion models. Diffusion Models [19] are generative models that gradually add noise to data through a forward diffusion process, followed by a reverse denoising process that reconstructs the original sample. The forward process corrupts the data sample x_0 through iterative noise addition controlled by a schedule α_t:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\alpha_t}\, x_{t-1}, \alpha_t I), \tag{1}$$

where α_t is a predefined noise schedule. The noisy sample x_t can be computed by $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$, where $\bar{\alpha}_t = \prod_{i=1}^{t}(1-\alpha_i)$ and $\epsilon \in \mathcal{N}(0,1)$ is a random Gaussian noise. The reverse process, parameterized by a neural network p_θ, learns a time-conditioned model to remove the noise added at each step.

Latent Diffusion Models (LDMs) work in a compressed latent space z_t rather than the high-dimensional data space [37], where the data x_0 is encoded as z_0 through an encoder. LDMs learn to minimize the objective,

$$L = \mathbb{E}_{z_0,t,\epsilon\sim\mathcal{N}(0,1)}\left[\|\epsilon - \epsilon_\theta(z_t, t)\|_2^2\right], \tag{2}$$

where ϵ_θ is the model's predicted noise at time t. Owing to the relevance to score-based generation models, the model estimates the log density of the distribution z_t, i.e., $\epsilon_\theta(z_t, t) \approx -\nabla_{z_t} \log p(z_t)$.

Classifier-free guidance. In the denoising process, different conditional inputs c (text, image, depth, mask, etc.) can be added to control the generation, so the denoising model predicts $\epsilon_\theta(z_t, t, c)$. In classifier-free guidance [18], a single neural network is used to parameterize both the unconditional denoising diffusion model $p_\theta(z)$ and conditional denoising diffusion model $p_\theta(z|c)$. While training, with some probability p_{uncond}, the unconditional model receives a null token, $\varnothing$ as c. p_{uncond} is set as a hyperparameter. While sampling, a linear combination of conditional and unconditional score estimates is used:

$$\tilde{\epsilon}_\theta(z_t, t, c) = (1 + w)\epsilon_\theta(z_t, t, c) - w\epsilon_\theta(z_t, t). \quad (3)$$

3.2. PathDiff

Histopathology image synthesis requires both high-level semantic guidance and precise structural fidelity. Existing methods [47] often require paired text and mask data, limiting their practicality due to the scarcity of such paired datasets in the histopathology domain. To address these challenges, we propose a unified framework, PathDiff, that leverages unpaired text and mask conditions, enabling fine-grained control over semantic and spatial features. This approach eliminates the need for paired data, supporting flexible and realistic image generation for downstream tasks.

Formally, let $D_{\text{T2I}} = \{(x_t, c_t)_i\}$ and $D_{\text{M2I}} = \{(x_m, c_m)_i\}$ represent two unpaired datasets, where D_{T2I} consists of image-text pairs and D_{M2I} contains image-mask pairs, with no overlapping images between the two datasets. We aim to learn a latent denoising diffusion model $p_\theta(z_t|t, c_t, c_m)$ that generates image samples conditioned on text c_t and mask c_m.

Joint training on unpaired datasets. The training pipeline, illustrated in Fig. 2 (a) and detailed in Algorithm 1, takes as input triplets of noisy latent, conditional mask, and conditional text, (x_t, c_m, c_t). We adopt a sampling strategy alternating between T2I and M2I datasets, allowing the model to learn from both sources and effectively integrate text and mask conditions. Specifically, we jointly train PathDiff by sampling data from the T2I dataset with probability p_{split} and from the M2I dataset with probability $1 - p_{\text{split}}$. When sampling from D_{T2I} dataset, we set $c_m = \varnothing_m$, leading to a training triplet of $(x_t, \varnothing_m, c_t)$. Similarly, when sampling from D_{M2I}, we set $c_t = \varnothing_t$, resulting in a training sample of $(x_m, c_m, \varnothing_t)$. Following existing methods [37], we set $\varnothing_t$ as an empty string. For $\varnothing_m$, while previous works [31, 32, 44] use a zero vector, we instead use a mask filled with an invalid label, as zero in our dataset indicates a background mask. Additionally, for robustness, we apply unconditional training by setting both $c_t = \varnothing_t$ and $c_m = \varnothing_m$ with a small probability p_{uncond}, as suggested in [47]. This training pipeline effectively allows PathDiff to learn from both unpaired conditions, enhancing its ability to generate realistic images conditioned on both modalities.

Algorithm 1 Joint Training on Unpaired Datasets

Require: p_{uncond}: probability of unconditional training, Text-to-Image dataset $\mathbf{D_{T2I}}$, Mask-to-Image dataset $\mathbf{M_{T2I}}$, p_{split}: probability of sampling from $\mathbf{D_{T2I}}$

repeat
  if $\mathbf{u} \sim \text{Uniform}[\mathbf{0}, \mathbf{1}] \geq \mathbf{p}_{\text{split}}$ **then**
    $(\mathbf{z_0}, \mathbf{c_m} = \varnothing_\mathbf{m}, \mathbf{c_t}) \sim \mathbf{p_1}(\mathbf{z_0}, \mathbf{c_t})$ ▷ Sample data from D_{T2I}
    $\mathbf{c_t} \leftarrow \varnothing_\mathbf{t}$ with probability $\mathbf{p}_{\text{uncond}}$ ▷ Randomly discard text
  else
    $(\mathbf{z_0}, \mathbf{c_m}, \mathbf{c_t} = \varnothing_\mathbf{t}) \sim \mathbf{p_2}(\mathbf{z_0}, \mathbf{c_m})$ ▷ Sample data from D_{M2I}
    $\mathbf{c_m} \leftarrow \varnothing_\mathbf{m}$ with probability $\mathbf{p}_{\text{uncond}}$ ▷ Randomly discard mask
  $\mathbf{t} \sim \text{Uniform}\{\mathbf{1}, .., \mathbf{T}\}$ ▷ Sample timestep
  $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ ▷ Sample noise
  $\mathbf{z_t} = \sqrt{\bar{\alpha}_\mathbf{t}}\,\mathbf{z_0} + \sqrt{\mathbf{1} - \bar{\alpha}_\mathbf{t}}\,\epsilon$ ▷ Corrupt data
  Take gradient step on:
    $\|\epsilon - \epsilon_\theta(\mathbf{z_t}, \mathbf{t}, \mathbf{c_m}, \mathbf{c_t})\|^\mathbf{2}$ with respect to ∇_θ ▷ Optimizing PathDiff
until converged

Following ControlNet [47] to add spatial control over image generation, we duplicate parts of the U-Net's downsampling and middle layers, adding *zero convolution* layers to these duplicates. Outputs from these layers are integrated into the original U-Net's skip connections. U-Net was pretrained on the T2I histopathology dataset before duplicating. Conditional mask is embedded with shallow CNN before adding as input to the U-Net encoder, while text embeddings are extracted from PLIP: pathology CLIP [21] and crossattended with U-Net layers. We use VAE from [45] trained on the TCGA-BRCA [9] histopathology dataset to encode and decode images. However, as noted by [45], reconstruction loss significantly affects the generation quality; therefore, we study its effect in supplementary. As shown in Fig. 2, all models are frozen except the copied-U-Net encoder.

We employ a latent diffusion pipeline [37] to train PathDiff. A shared VAE encoder-decoder is used for both datasets, resulting in a unified latent representation z. This approach assumes that the VAE can compress and reconstruct images from both datasets without significant loss, maintaining a consistent latent representation across domains.

Optimization loss. PathDiff jointly optimizes the diffusion model p_θ to predict the noise added at every step t of the noising process. When the training sample is sampled from D_{T2I} it minimizes loss with respect to sample $x_t \in D_{\text{T2I}}$,

$$L_{\text{T2I}} = \mathbb{E}_{z_0, t, \epsilon \sim \mathcal{N}(0,1)} \left[\|\epsilon - \epsilon_\theta(z_t, t, \varnothing_m, c_t)\|_2^2 \right] \quad (4)$$

similarly, for training samples from D_{M2I} it minimizes the loss with respect to sample $x_m \in D_{M2I}$,

$$L_{\text{M2I}} = \mathbb{E}_{z_0, t, \epsilon \sim \mathcal{N}(0,1)} \left[\|\epsilon - \epsilon_\theta(z_t, t, c_m, \varnothing_t)\|_2^2 \right] \quad (5)$$

Unified conditional sampling. Since PathDiff is jointly trained on both D_{T2I} and D_{M2I} datasets, it can be queried to sample images conditioned on both c_t and c_m, effectively unifying these conditions as guidance, as illustrated

in Fig. 4 (c). The image generation process is illustrated in Fig. 2 (b). During inference, we generate images using classifier-free guidance [18], updating the predicted noise with the model using Eq. (3). The conditioning variable c can take values from $\mathbf{c} \in \{(\varnothing_{\mathbf{m}}, \mathbf{c_t}), (\mathbf{c_m}, \varnothing_{\mathbf{t}}), (\mathbf{c_m}, \mathbf{c_t})\}$. We can selectively generate images from either distribution by setting one of the conditions to $\varnothing$. Specifically,

$$z_t \sim \begin{cases} p_\theta(z_t|t, c_t) & \text{if } c = (\varnothing_m, c_t), \\ p_\theta(z_t|t, c_m) & \text{if } c = (c_m, \varnothing_t), \\ p_\theta(z_t|t, c_m, c_t) & \text{if } c = (c_m, c_t). \end{cases}$$

Tab. 1 concretely defines these variations. Our model can selectively generate images based on the conditions illustrated in Fig. 3 and Fig. 4 (b). We present extensive experiments on images drawn from each model variant in Sec. 4.

Method	Training		Generation	
	Mask	Text	Mask	Text
Diffmix [31]	✓	✗	✓	✗
SDM [44]	✓	✗	✓	✗
ControlNet-M [47] [a]	✓	✗	✓	✗
ControlNet-T [47] [a]	✓	✗	✗	✓
Ours-M	✓	✓	✓	✗
Ours-T	✓	✓	✗	✓
Ours-M+T	✓	✓	✓	✓

Table 1. **Conditioning inputs at training and generation stages for all methods**. Method-**T** uses **T**ext for generation, Method-**M** uses **M**ask for generation, Method-**M+T** uses **B**oth Text and Mask for generation. [a] ControlNet-M and ControlNet-T are not trained on the T2I dataset PathCap [40]. However, it uses locked Stable Diffusion [37] backbone finetuned on PathCap for training and generation.

4. Experiments

We evaluate our method in two main aspects: the quality of the generated images and the effectiveness of using synthetic images as additional training data for downstream segmentation tasks. We evaluate image generation quality in Sec. 4.3, focusing on image fidelity, image-mask faithfulness, and alignment quality between images and text. Additionally, the utility of the synthetic images in downstream segmentation tasks is detailed in Sec. 4.4. We also discusses the domain experts' survey results in Sec. 4.5. As outlined in Tab. 1, PathDiff was jointly trained on M2I and T2I datasets; however, we can generate images conditioned on text, mask, or both. Each of these variants is evaluated separately. Similarly, we use two variants of ControlNet [47] (as defined in Tab. 1), for a fair and inclusive comparison.

4.1. Datasets

We use three mask-image histopathology datasets (Panuke [12], CoNIC [15], and MoNuSAC [43]) and one text-image dataset (PathCap [40]) to conduct a comparative analysis across multiple metrics, evaluating various diffusion methods and their related downstream tasks.

PanNuke dataset [12] provides annotated histopathology images across 19 tissue types with 189,744 annotated nuclei in five classes (neoplastic, inflammatory, connective, dead, and epithelial). The dataset is highly imbalanced [23] and one of the most challenging to perform the segmentation task [25].

CoNIC dataset [15] is one of the most extensive publicly available histopathology datasets, containing approximately 535,000 labeled colon nuclei across six cell types: epithelial, lymphocytes, plasma cells, eosinophils, neutrophils, and connective tissue cells.

MoNuSAC [43] spans four organs (lung, prostate, kidney, and breast) and includes over 46,000 annotated nuclear boundaries collected from diverse sources for four cell types: epithelial cells, lymphocytes, macrophages, and neutrophils.

PathCap [40] includes 207K pathology image-caption pairs, 197K from PubMed and guidelines, and 10K annotated by expert cytologists in liquid-based cytology (LBC). We used a 100K subset of PathCap containing only H&E-stained histology images to ensure consistency with the three mask-image datasets, which are also H&E-stained.

Data splits: All the four datasets were split in an 8:2 train-to-test ratio. Train and test split for downstream tasks coincides with the split of generation experiments. For downstream tasks, we use synthetic samples generated exclusively from the training split as additional training data in all experiments, ensuring a fair comparison among the synthetic models.

4.2. Implementation Details

PathDiff is trained on 256×256 image patches. We resize patches for the MoNuSAC [43] and PathCap [40] datasets by taking crops of size 256×256 from WSIs. PanNuke [12] and CoNIC [15] datasets both have images with size 256×256. Our model is trained for 60 epochs on 4 NVIDIA A6000 GPUs with a batch size 72 and a learning rate of 3.75×10^{-5} with 1000 warmup steps. We use DDIM sampling [38] with 100 steps and a classifier-free guidance scale set to 1.5 for image generation. We train CellViT on a single NVIDIA RTX A6000 for 130 epochs (with the first 30 epochs frozen), using a batch size of 16 and a learning rate of 0.001, retaining all other training parameters from the original work.

4.3. Image Generation Quality

Image Generation Fidelity: To evaluate the quality of generated images, we use CLIP-FID following [16, 34]. As FID

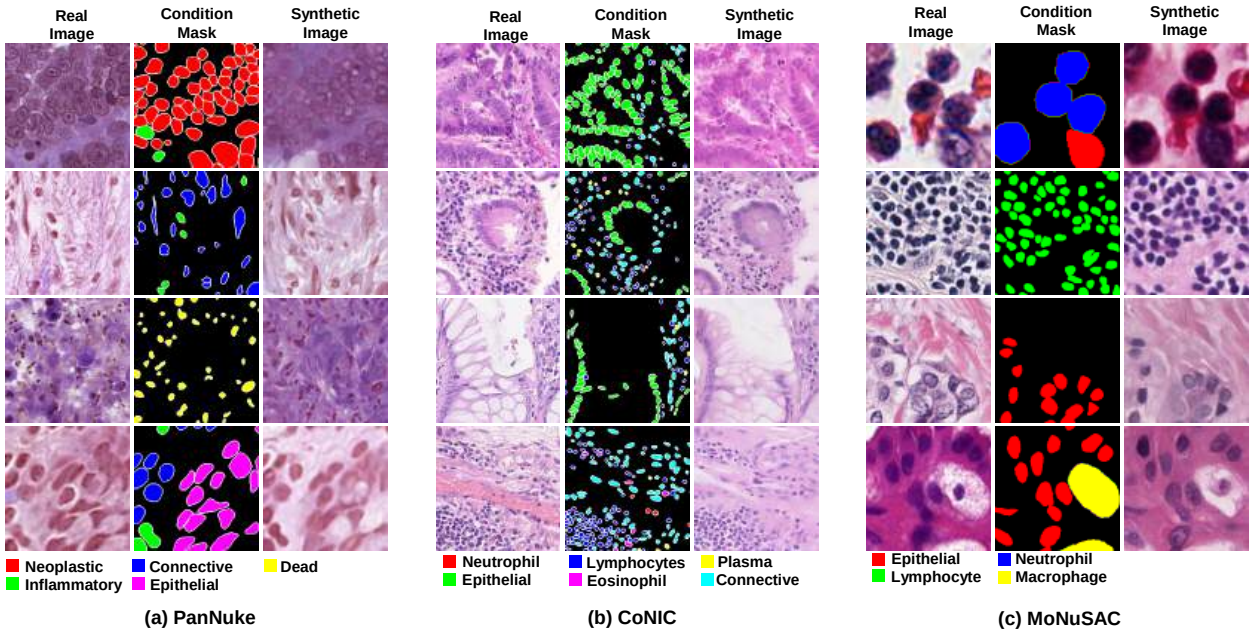


Figure 3. **Synthetic images generated by PathDiff** on Mask to image datasets (a) PanNuke [12], (b) CoNIC [15], (c) MoNuSAC [43]. PathDiff closely follows the spatial structures defined by the cell type label map, producing synthetic images that closely resemble the spatial arrangement of real images.

Method	PanNuke				CoNIC				MoNuSAC			
	FID ↓	KID ↓	FID ↓	KID ↓	FID ↓	KID ↓	FID ↓	KID ↓	FID ↓	KID ↓	FID ↓	KID ↓
Diffmix [31]	7.28	0.0492	8.32	0.0536	8.09	0.0966	8.58	0.0937	9.25	0.1036	16.78	0.35
SDM [44]	7.14	0.1147	**7.13**	0.1149	8.27	0.0709	9.80	0.0737	7.46	0.1278	9.44	0.1340
ControlNet-M [47]	16.44	0.0705	15.36	0.7722	6.76	0.0790	6.79	0.0695	20.06	0.1050	20.56	0.1068
Ours-M	**6.94**	**0.0389**	7.21	**0.0415**	**5.64**	**0.0524**	**5.54**	**0.0488**	**6.71**	**0.0616**	6.99	**0.0758**
Ours-M+T	11.03	0.0718	12.32	0.0952	6.28	0.0730	6.73	0.0758	7.16	0.0782	**6.60**	0.0874

Table 2. **Comparison of CLIP-FID [17, 36] and KID [2] across training and test splits for PanNuke [12], CoNIC [15], and MoNuSAC [43].** Ours-M and Ours-M+T are trained jointly with the text-image dataset: PathCap [40]. The best results are in **bold** and the second best is underlined.

Method	PanNuke		CoNIC		MoNuSAC	
	FS1 ↑	FS2 ↑	FS1 ↑	FS2 ↑	FS1 ↑	FS2 ↑
SDM [44]	0.7025	0.6826	0.7433	0.7376	0.6377	0.6104
ControlNet-M [47]	0.6990	0.6821	0.7629	0.7537	0.6317	0.6081
Diffmix [31]	0.7419	0.6964	0.7597	0.7163	0.6488	0.6182
Ours-M	**0.7437**	**0.7406**	**0.7873**	**0.7632**	**0.6519**	**0.6308**
Ours-M+T	0.7324	0.7187	0.7581	0.7245	0.6412	0.6238

Table 3. **Faithfulness scores for the PanNuke [12], CoNIC [15], and MoNuSAC [43] datasets.** FS1 measures the adherence of synthetic images to the spatial structure of real masks by computing DICE(M_{syn}^{pred}, M_{real}). At the same time, FS2 assesses the similarity between synthetic and real images for potential downstream tasks, calculated as DICE($M_{syn}^{pred}, M_{real}^{pred}$).

is sensitive to the number of images, we also report Kernel Inception Distance (KID) [2] to provide an unbiased estimate. As shown in Tab. 2, PathDiff-M achieves the **best** generation quality in terms of FID and KID compared against the existing mask-to-image generation methods: Diffmix [31], SDM [44] and ControlNet [47]. Fig. 3 presents qualitative examples generated by PathDiff-M.

Mask-to-Image Faithfulness: To evaluate whether the generated images adhere to the spatial guidance provided by the masks, we use the evaluation protocol of SPADE [33], also termed as Faithfulness Score (FS) [26]. Auxiliary segmentation model [23] S is trained on train split of mask-image datasets. Real images I_{real} from test split (I_{real}, M_{real}) are passed through S to get predicted segmentation masks M_{real}^{pred}. Synthetic images generated by

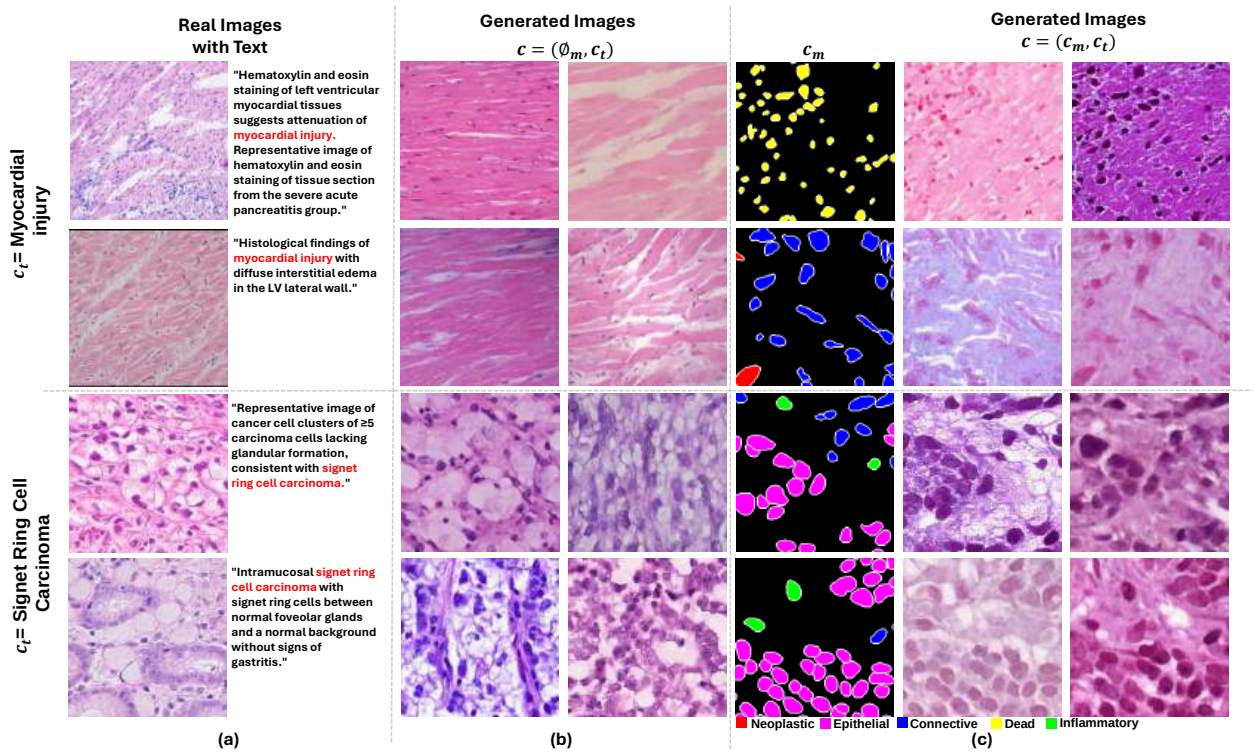


Figure 4. **Qualitative examples demonstrating the effectiveness of unifying guidance from both text and mask conditions for image generation.** (a) Real images are shown with their respective text-based descriptions. (b) PathDiff-T generated images conditioned solely on text replicate the visual attributes observed in real images, with the mask condition set to $\varnothing_m$. (c) PathDiff-M+T generated images conditioned on the same text c_t and an additional mask c_m accurately incorporate the visual features of the real images and adhere to the specified cell mask structure.

PathDiff-M conditioned on masks M_{real} are passed through S to get predicted segmentation masks M_{syn}^{pred}. FS1 calculates DICE$(M_{syn}^{pred}, M_{real})$, FS2 on other hand calculates DICE$(M_{syn}^{pred}, M_{real}^{pred})$. As shown in Tab. 3, we attain the best FS1 and FS2 among methods, indicating that the synthetic images closely align with the real masks and reach a similarity to real images that enhances their utility for downstream segmentation tasks. PathDiff-M+T, conditioned on text and mask, achieves competitive FS1 and FS2 scores, demonstrating its ability to incorporate image features relevant to text without compromising mask alignment. PathDiff-M intuitively achieves the **highest score** for image-mask alignment compared to PathDiff-M+T. Notably, compared with the counterparts [31, 44, 47], which also sample from the single mask condition, the joint training strategy in PathDiff-M demonstrates a superior ability to learn from the limited data.

Text-to-Image Alignment: Fig. 4 presents qualitative samples generated by PatDiff-T conditioned on only text; images generated replicate visual attributes about the text as observed in real images. Tab. 4 presents FID and KID scores compared to ControlNet-T [47] conditioned only on text. We also use metric - **PLIP** [21] **cosine similarity** to evaluate the consistency between image and text embeddings. High PLIP image similarity scores indicate better alignment between text and images. Ours-T has the highest PLIP score of $\geq$ **24.17**, indicating good alignment between texts and generated images. Ours-T variant performs better than ControlNet-T on FID [17] and KID [2] across all M2I constituent datasets, indicating large T2I datasets can be jointly trained with a wide range of M2I datasets.

4.4. Downstream tasks

To evaluate the utility of the generated images, we use them as additional training data for segmentation model: CellViT [20] (SAM-B variant) backbone with HoVer-Net [14] as the decoder. Tab. 5 compares segmentation and classification performance across multiple metrics, including Dice, Jaccard, AJI (Aggregated Jaccard Index), HD95 [22] (95th percentile Hausdorff Distance), F1 score, Precision, and Recall. We note that our proposed method consistently improves the performace across various metrics on all the three datasets. In segmentation tasks on PanNuke, Ours-M achieves the highest Dice score **(0.8164)** and performs well across other metrics like Jaccard (0.7122) and AJI (0.7117). Similarly, in

Method	M2I Dataset: PanNuke						M2I Dataset: CoNIC						M2I Dataset: MoNuSAC					
	Train			Test			Train			Test			Train			Test		
	FID ↓	KID ↓	PLIP ↑	FID ↓	KID ↓	PLIP ↑	FID ↓	KID ↓	PLIP ↑	FID ↓	KID ↓	PLIP ↑	FID ↓	KID ↓	PLIP ↑	FID ↓	KID ↓	PLIP ↑
ControlNet-T [47]	45.39	0.0872	23.27	45.04	0.0886	22.95	26.32	0.1275	21.97	25.10	0.1269	21.98	28.66	0.1111	21.88	29.86	0.1091	21.76
Ours-T	18.52	**0.0619**	**24.18**	19.60	**0.0644**	**24.05**	18.97	**0.0640**	**24.03**	19.96	**0.0655**	**24.17**	13.72	**0.0538**	**24.17**	**14.04**	**0.0557**	**24.66**
Ours-M+T	**14.39**	0.0884	23.01	**14.26**	0.1059	22.80	**11.12**	0.1059	22.09	**11.78**	0.1032	21.75	**12.64**	0.1338	22.18	14.19	0.1764	21.66

Table 4. **Comparison of CLIP-FID [17, 36], KID [2], and PLIP [21]similarity scores for training and test splits for text-image dataset: PathCap**. Ours-M and Ours-M+T are trained jointly with three mask-image datasets: PanNuke [12], CoNIC [15], and MoNuSAC [43]. PLIP similarity scores on real PathCap [40] train and test splits are **26.34** and **26.56**, respectively, as a reference for the comparison.

classification tasks on PanNuke, Ours-M achieves a strong F1 score of **0.8161** with balanced precision and recall. On the CoNIC dataset, Ours-M again leads in segmentation with a Dice score of **0.8356** and performs well in classification with an F1 score of 0.8053. On MoNuSAC, where performance is generally lower across all methods, Ours-M still outperforms others in segmentation with a Dice score of **0.7221** and achieves an F1 score of **0.7115** for classification. Notably, Ours-M+T, and in some cases the baseline, achieve the second-best performance on downstream tasks. Pertinent to mention is that when evaluating the classification and segmentation performance, the split used included only real test data.

4.5. Domain expert assessment

As pointed out in [1], we acknowledge that traditional fidelity metrics like FID [17] are only somewhat applicable to histological images as large image datasets like ImageNet [5] would unlikely have images from this specific domain. Therefore, we conduct expert evaluation to validate the efficiency of the generated samples. We surveyed two domain experts—a physician and a pathology researcher—to review the generated data, to assess if the samples accurately reflect the characteristics of real specimens.

Expert Image Preference Assessment: We presented domain experts with 200 synthetic images generated from PathDiff, SDM [44], ControlNet [47], DiffMix [31]. Each Quadruplet of images was generated using the same conditional mask. Domain experts were asked to choose one of the four images that looked most real. As shown in Fig. 5 (a), domain experts **preferred PathDiff-generated images** significantly more than the existing SOTA methods, indicating our generated images look more realistic to an expert eye.

Expert Turing Test: In this experiment, domain experts are presented with an equal number of real and synthetic images in random order. Real labels of images are hidden. We ask to choose whether the given image looks *real* or *synthetic*. As presented in the confusion matrix in Fig. 5 (b), domain experts struggled to consistently identify real and synthetic images, with many misclassifications occurring in both directions (real images labeled as synthetic and vice versa). Of all synthetic images presented, **31** were misclassified as real by domain experts, compared to **only 21** correctly identified as synthetic. The high misclassification rate of synthetic images (approximately 60% labeled as real) indicates that our synthetic images are realistic enough, making it difficult for domain experts to distinguish them from real images.

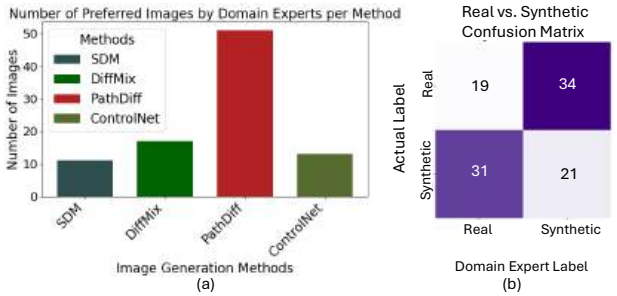


Figure 5. **Results of Domain expert survey:** (a) PathDiff generated images are significantly preferred as *real* by experts. (b) Roughly 60% of *synthetic* images (31 out of 52) were labeled as real, suggesting *synthetic* images looked real to an expert eye.

5. Conclusion

To address data scarcity in histopathology image analysis, we propose a novel diffusion framework that simultaneously leverages diagnostic reports as contextual guidance and mask inputs for precise spatial control. We demonstrate that a joint conditional diffusion model can be learned to unify conditions from unpaired data using classifier-free guidance maximizing the use of limited existing data and allowing us to achieve the best performance across a variety of experiments evaluating image quality and downstream tasks.

Acknowledgments

I am deeply grateful to all co-authors for their invaluable contributions and collaborative spirit throughout this project. I extend my sincere thanks to Mahesh Bhosale, Yuanhao Zhai, Yunjie Tian, and Nan Xi at the University at Buffalo; Samuel Border and Pinaki Sarder at the University of Florida; and Xuan Gong at Harvard Medical School. Their expertise, dedication, and teamwork were instrumental in making this work possible. I am especially thankful to my advisor, Prof.

Dataset	Method	Segmentation				Classification		
		Dice ↑	Jaccard ↑	AJI ↑	HD (95) ↓	F1 ↑	Precision ↑	Recall ↑
PanNuke	Baseline	0.7834	0.6951	0.6904	6.81	0.7826	0.7472	0.8215
	w/ SDM	0.7891	0.6816	0.6729	6.26	0.7792	0.8014	0.7581
	w/ ControlNet	0.8093	0.6979	0.6954	5.87	0.8134	**0.8354**	0.7925
	w/ Ours-M	**0.8164**	**0.7122**	**0.7117**	5.12	**0.8161**	0.7827	**0.8524**
	w/ Ours-M+T	0.7947	0.7040	0.7018	5.74	0.7949	0.7803	0.8100
CoNIC	Baseline	0.7927	0.6978	0.6917	4.38	0.7614	0.7153	0.8138
	w/ SDM	0.8007	0.6846	0.6846	4.50	0.7883	0.7620	0.8164
	w/ ControlNet	0.8076	0.6999	0.6950	3.16	0.8000	**0.8344**	0.7683
	w/ Ours-M	**0.8356**	**0.7195**	**0.7141**	**2.97**	**0.8053**	0.7795	**0.8328**
	w/ Ours-M+T	0.8114	0.7003	0.7003	3.41	0.7751	0.7572	0.7938
MoNuSAC	Baseline	0.7089	0.6131	0.6097	6.77	0.6652	0.6383	0.6944
	w/ SDM	0.6783	0.5744	0.5731	6.49	0.6685	0.6940	0.6448
	w/ ControlNet	0.6940	0.5852	0.5800	6.54	0.6843	0.6533	0.7183
	w/ Ours-M	**0.7221**	**0.6197**	**0.6192**	**5.88**	**0.7115**	**0.6951**	**0.7286**
	w/ Ours-M+T	0.7124	0.6072	0.6067	6.25	0.6861	0.6569	0.7187

Table 5. **Comparison of segmentation and classification metrics across different datasets and methods on CellViT model**. The Baseline is only trained with real data, and the rest of the methods have synthetic data added in equal proportion to the train set.

David Doermann, for his unwavering guidance, mentorship,
and support throughout this research.

PathDiff: A Diffusion Framework for Histopathology Image Synthesis Unifying Unpaired Text and Mask Conditions

Technical Report

Supplementary Material

6. Scaling Augmentation in Downstream Tasks

To evaluate the impact of synthetic data augmentation on downstream tasks, we created three augmented sets for training the CellViT [23] model on the PanNuke [12] dataset. These augmented datasets were generated from the training split, conditioned on both masks and text descriptions, and evaluated on the real test split. They were incrementally added to the training process while keeping the size of the real train split constant.

Scaling the augmentation set progressively improves the classification and segmentation performance of the training data, as shown in Fig. 6, with the 3x and 2.5x augmented synthetic datasets outperforming the relatively smaller ones on all metrics. After 2.5x, the performance metrics plateau. These results demonstrate that PathDiff effectively contributes valuable synthetic data in every augmented set, consistently improving model performance across all downstream tasks as the size of the synthetic set increases, highlighting the utility of PathDiff in generating high-quality data for histopathology image analysis.

7. Qualitative Comparison of Synthetic Images

In this section, we present a qualitative comparison of synthesized images generated by PathDiff, DiffMix [31], SDM [44], and ControlNet [47].

7.1. Mask-to-Image examples

Fig. 7 shows a comparison of synthetic images generated on the PanNuke [12], CoNIC [15], and MoNuSAC [43] datasets. As illustrated in the figure, images generated by DiffMix appear very coarse with additional artifacts and fail to preserve the accurate stain colors observed in the original images. Consistent with observations reported by [28], we find that SDM-generated images display unrealistic color overlay artifacts. The color distribution of ControlNet-generated images appears highly inconsistent, being significantly inaccurate in some cases while better than others in certain instances. On the other hand, PathDiff accurately follows the cell mask and maintains the stain colors.

7.2. Text-to-Image examples

Fig. 8 shows samples generated by PathDiff and ControlNet. As with images conditioned on masks, ControlNet fails to preserve details in the original image and exhibits impractical colors uncommon in histopathology images. This explains the high FID and KID values compared to PathDiff in Tab.3 of the main paper.

8. Considerations for choosing values of p_{split}

When training jointly on two datasets—Text-to-Image and Mask-to-Image—p_{split} controls the proportion of data sampled from each of them. We evaluate performance with three values of p_{split}: 0.2, 0.5, and 0.8. Results using only text are shown in Table 6, mask-only conditioning in Tab. 7, and both text and mask conditioning in Tab. 8.

In these experiments, $p_{split} = 0.5$ strikes a balance, explaining why we chose this value in the main paper. While it seems logical to assign a larger probability to the larger dataset to cover more of its samples, we found that $p_{split} = 0.5$ works well in practice.

p_{split}	PathCap: Train			PathCap: Test		
	FID ↓	KID ↓	PLIP ↑	FID ↓	KID ↓	PLIP ↑
$p_{split} = 0.2$	16.33	0.0624	24.43	15.74	0.0603	24.50
$p_{split} = 0.5$	18.52	0.0619	24.18	19.60	0.0644	24.05
$p_{split} = 0.8$	19.58	0.0649	24.34	18.87	0.0626	24.27

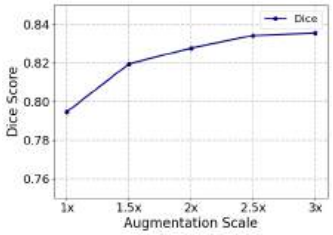
Table 6. **Considerations for p_{split}.** CLIP-FID [17, 36], KID [2], and PLIP [21] similarity scores for different p_{split} values on PathCap [40], with text condition c_t used for sampling. PLIP [21] similarity scores on the real PathCap train and test splits are **26.34** and **26.56**, respectively, provided as a reference for comparison.

p_{split}	PanNuke: Train		PanNuke: Test	
	FID ↓	KID ↓	FID ↓	KID ↓
$p_{split} = 0.2$	7.36	0.0525	7.88	0.0559
$p_{split} = 0.5$	6.94	0.0389	7.28	0.0415
$p_{split} = 0.8$	8.57	0.0584	8.97	0.0707

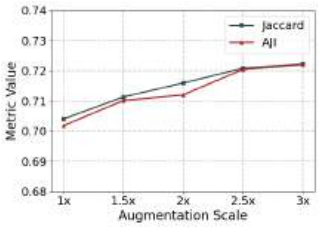
Table 7. **Considerations for p_{split}.** CLIP-FID [17, 36], KID [2] for different p_{split} values on PanNuke [12] dataset. Only mask condition c_m was used for sampling.

9. Finetuning VAE

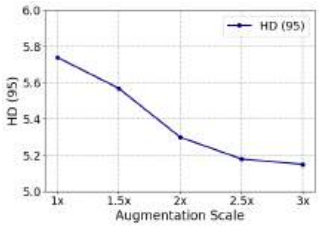
The reconstruction performance of VAEs [37, 45] plays a crucial role in the fidelity of generated images. Losses intro-



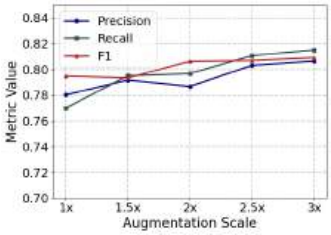
(a) Segmentation - Dice



(b) Segmentation - Jaccard & AJI



(c) Segmentation - HD (95)



(d) Classification

Figure 6. **Comparison of segmentation and classification metrics on the PanNuke [12] dataset across augmentation scaling factors.** The addition of PathDIff-generated synthetic data consistently increases downstream classification and segmentation performance. 1x uses one *synthetic* augmentation set equal to the *real* train split size; 1.5x adds another *synthetic* set equal to 1.5 times the *real* train split size and so on. After 2.5x, the performance metrics plateau.

p_{split}	PanNuke				PathCap					
	Train		Test		Train			Test		
	FID ↓	KID ↓	FID ↓	KID ↓	FID ↓	KID ↓	PLIP ↑	FID ↓	KID ↓	PLIP ↑
$p_{split} = 0.2$	10.25	0.0672	11.99	0.0800	15.21	0.0846	23.01	16.07	0.0956	22.81
$p_{split} = 0.5$	11.03	0.0718	12.32	0.0952	14.39	0.0884	23.01	14.26	0.1059	22.80
$p_{split} = 0.8$	9.723	0.0729	10.37	0.0862	12.78	0.0955	22.97	12.53	0.1107	22.70

Table 8. **Consideration for p_{split}.** CLIP-FID [17, 36], KID [2], and PLIP [21] similarity scores for different p_{split} values for PanNuke [12] and PathCap [40]. We used both text c_t and mask c_m for sampling.

duced during the compression and decompression stages in VAEs compound with the denoising process losses in subsequent stages, directly impacting the quality of the generated images.

Initially, we used the VQ-VAE from [45], which was trained on the TCGA-BRCA [9] dataset containing whole-slide images (WSIs) exclusively from breast tissues. While this VAE outperforms the one from [37], which was trained on natural images, its applicability is limited as it lacks representation of diverse tissue types. We fine-tuned the VAE on the datasets used in this work, including PanNuke [12], PathCap [40], CoNIC [15], and MoNuSAC [43].

As demonstrated in Tab. 9, fine-tuning the VAE on these datasets results in improvements across all reconstruction and generation metrics. However, these improvements, while consistent, are relatively modest.

VAE Trained on	Metrics			
	LPIPS ↓	SSIM ↑	MSE ↓	FID ↓
TCGA-BRCA [9]	0.0462	0.7962	0.0084	6.94
Datasets: D	0.0429	0.8212	0.0070	6.31

Table 9. **Effect of fine-tuning VAE on datasets D:** PanNuke [12], PathCap [40], MoNuSAC [43], and CoNIC [15].

10. Details about Domain Expert Survey

We used an interactive web-based tool to conduct a domain experts survey. Clear instructions were given to evaluate the images. Fig. 9 and Fig. 10 show the web interface used for the domain expert image preference experiment and the Turing test respectively.

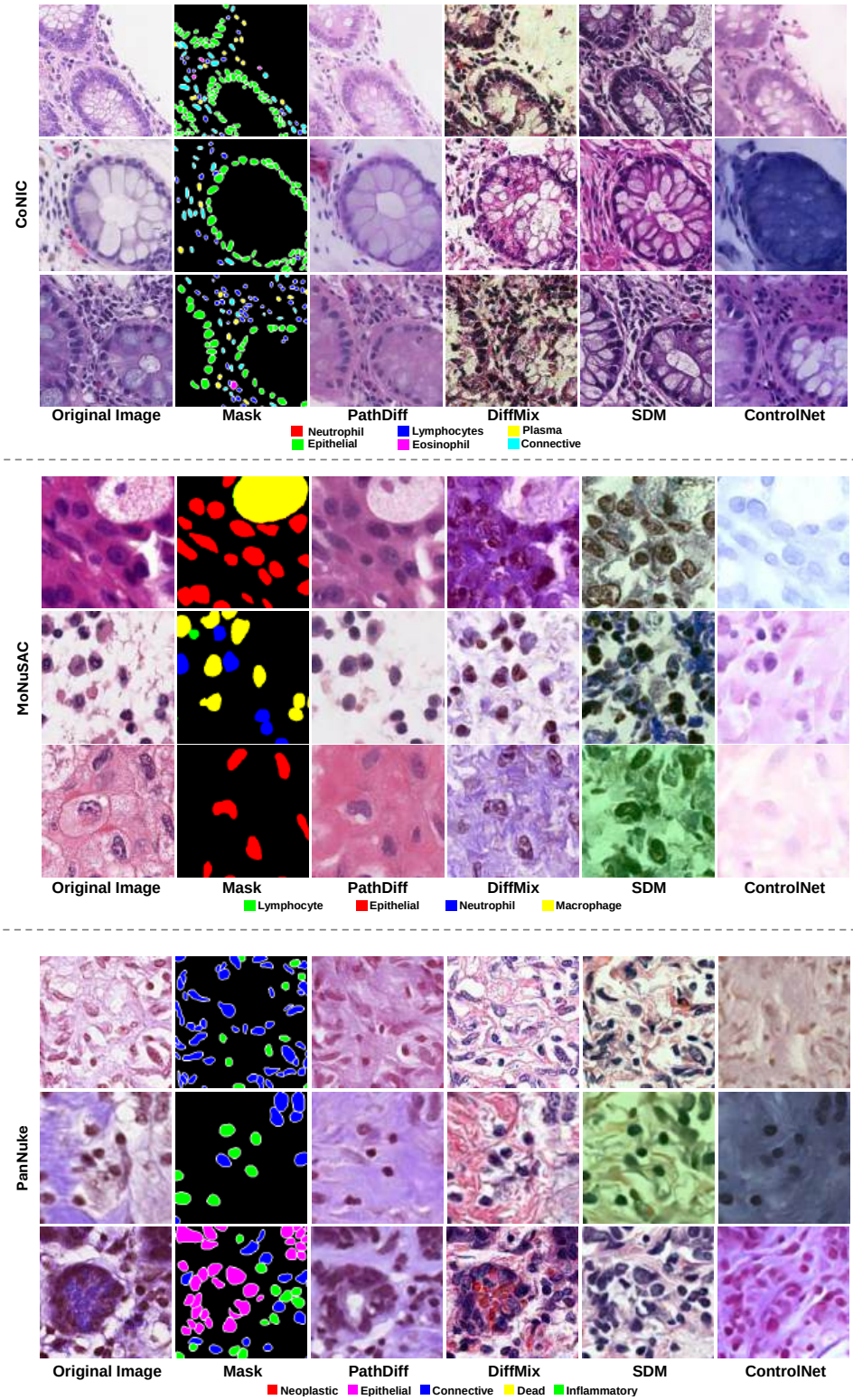


Figure 7. **Qualitative comparison** of synthetic images generated by PathDiff, DiffMix [31], SDM [44], and ControlNet [47] on the CoNIC [15], MoNuSAC [43], and PanNuke [12] datasets.

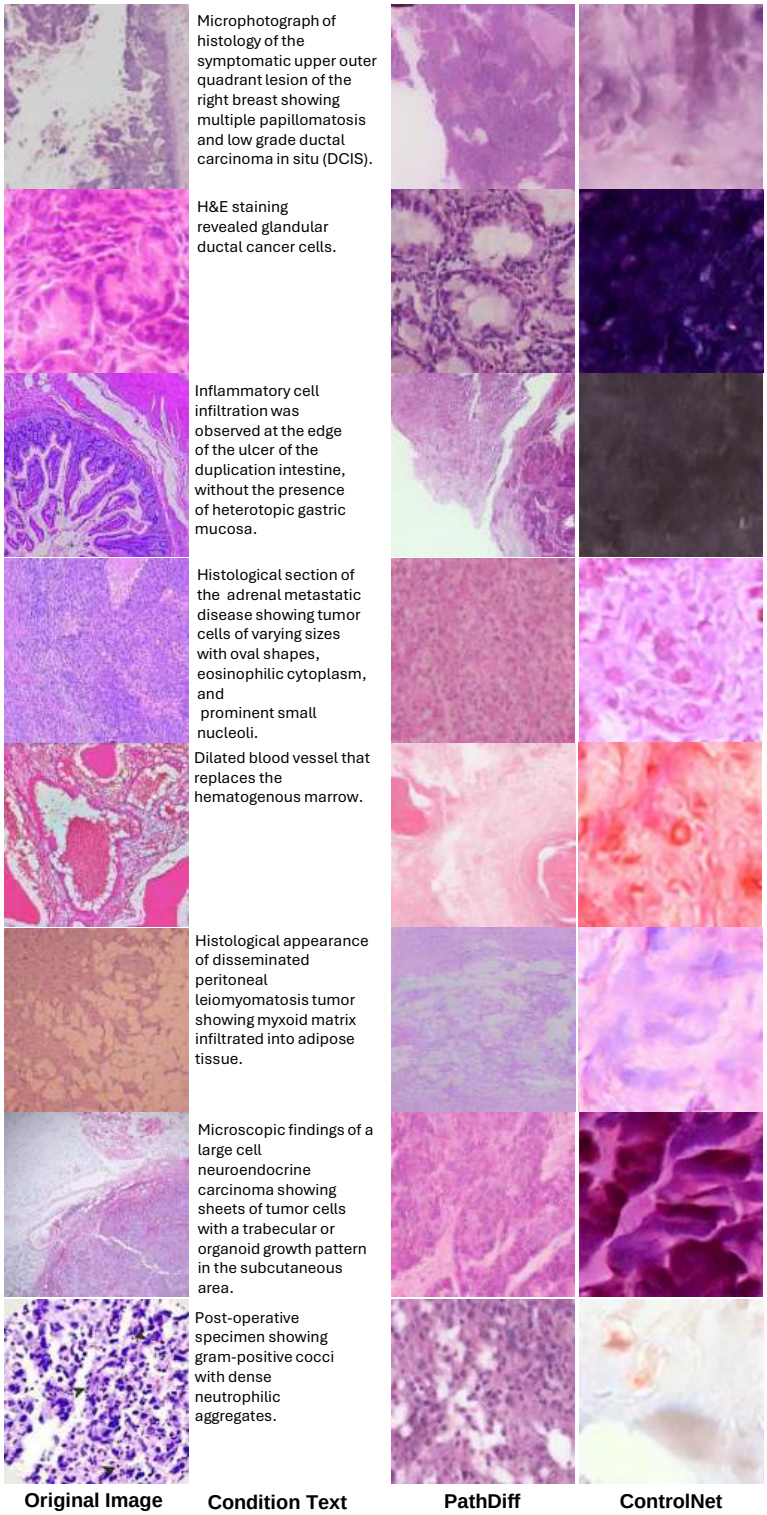


Figure 8. **Qualitative comparison** of synthetic images generated by PathDiff and ControlNet [47] on the PathCap [40] dataset.

11. Ethics Statement

This work aims to generate synthetic histopathology images to address data scarcity while ensuring patient privacy, and supporting research and algorithm development responsibly.

- **Privacy**: All data used is publicly available, ensuring no patient-identifiable information is involved. The generated synthetic images are entirely artificial and do not correspond to real patients.
- **Transparency**: We will open-source our code and models to promote transparency and enable evaluation and validation by the research community.
- **Responsible Usage**: The synthetic data is for research and training only and must not be used in clinical decision-making; Usage guidelines will be provided in the open-source repository for clarification.
- **Bias and Limitations**: Synthetic data may reflect biases inherent in the original datasets. We encourage users to critically evaluate these biases and use the data responsibly to avoid misuse and misrepresentation.

References

[1] Marco Aversa, Gabriel Nobis, Miriam Hägele, Kai Standvoss, Mihaela Chirica, Roderick Murray-Smith, Ahmed Alaa, Lukas Ruff, Daniela Ivanova, Wojciech Samek, Frederick Klauschen, Bruno Sanguinetti, and Luis Oala. 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. 2, 3, 8

[2] Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018. 6, 7, 8, 10, 11

[3] S. Butte, H. Wang, A. Vakanski, and M. Xian. Enhanced sharp-gan for histopathology image synthesis. In *Proceedings of IEEE International Symposium on Biomedical Imaging (ISBI)*, page 10.1109/isbi53787.2023.10230516, 2023. Epub 2023 Sep 1. 2

[4] Gagandeep B. Daroach, Savannah R. Duenweg, Michael Brehler, Allison K. Lowman, Kenneth A. Iczkowski, Kenneth M. Jacobsohn, Josiah A. Yoder, and Peter S. LaViolette. Prostate cancer histology synthesis using stylegan latent space annotation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 398–408, Cham, 2022. Springer Nature Switzerland. 2

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 8

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, pages 8780–8794. Curran Associates, Inc., 2021. 2

[7] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2024. Curran Associates Inc. 2

[8] James M. Dolezal, Rachelle Wolk, Hanna M. Hieromnimon, Frederick M. Howard, Andrew Srisuwananukorn, Dmitry Karpeyev, Siddhi Ramesh, Sara Kochanny, Jung Woo Kwon, Meghana Agni, et al. Deep learning generates synthetic cancer histology for explainability and education. *NPJ Precision Oncology*, 7(1):49, 2023. 2

[9] JN Cancer Genome Atlas Research Network et al. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120, 2013. 2, 4, 11

[10] Virginia Fernandez, Walter Hugo Lopez Pinaya, Pedro Borges, Petru-Daniel Tudosiu, Mark S. Graham, Tom Vercauteren, and M. Jorge Cardoso. Can segmentation models be trained with fully synthetically generated data? In *Simulation and Synthesis in Medical Imaging: 7th International Workshop, SASHIMI 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Proceedings*, pages 79–90. Springer, 2022. 2

[11] Michael Gadermayr and Maximilian Tschuchnig. Multiple instance learning for digital pathology: A review of the state-of-the-art, limitations & future potential. *Computerized Medical Imaging and Graphics*, 112:102337, 2024. 1

[12] Jevgenij Gamper, Navid Alemi Koohbanani, Ksenija Benet, Ali Khuram, and Nasir Rajpoot. Pannuke: An open pan-cancer histology dataset for nuclei instance segmentation and classification. In *Digital Pathology*, pages 11–19, Cham, 2019. Springer International Publishing. 5, 6, 8, 10, 11, 12

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[14] Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical image analysis*, 58:101563, 2019. 7

[15] Simon Graham, Quoc Dang Vu, Mostafa Jahanifar, Martin Weigert, Uwe Schmidt, Wenhua Zhang, Jun Zhang, Sen Yang, Jinxi Xiang, Xiyue Wang, et al. Conic challenge: Pushing the frontiers of nuclear detection, segmentation, classification and counting. *Medical image analysis*, 92:103047, 2024. 5, 6, 8, 10, 11, 12

[16] Alexandros Graikos, Srikar Yellapragada, Minh-Quan Le, Saarthak Kapse, Prateek Prasanna, Joel Saltz, and Dimitris Samaras. Learned representation-guided diffusion models for large-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8532–8542, 2024. 2, 5

[17] 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. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 6, 7, 8, 10, 11

[18] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. 2, 4, 5

Figure 9. **Interactive web interface** used for domain expert image preference experiment.

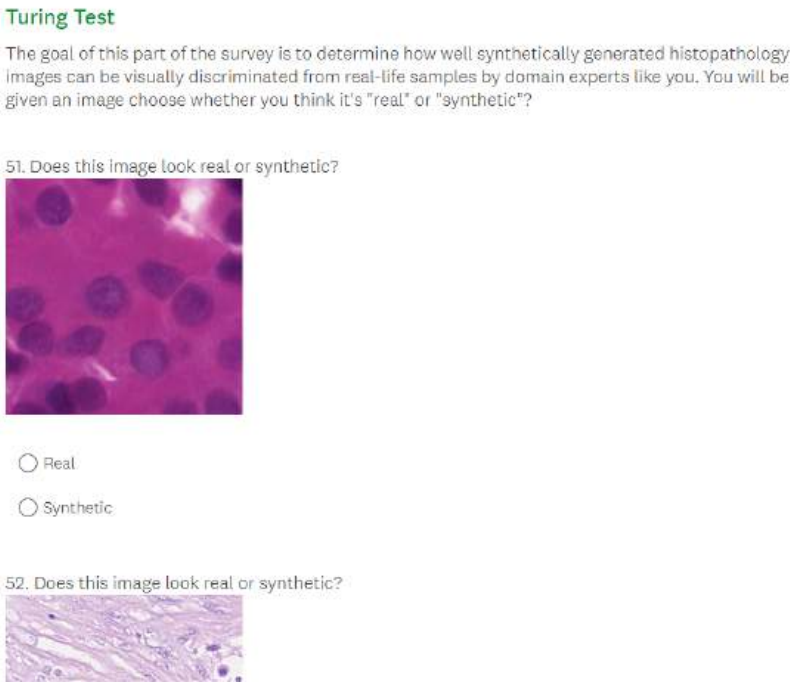


Figure 10. **Interactive web interface** used for domain expert Turing test.

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851, 2020. 2, 3

[20] Fabian Hörst, Moritz Rempe, Lukas Heine, Constantin Seibold, Julius Keyl, Giulia Baldini, Selma Ugurel, Jens Siveke, Barbara Grünwald, Jan Egger, et al. Cellvit: Vision transformers for precise cell segmentation and classification. *Medical Image Analysis*, 94:103143, 2024. 7

[21] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature Medicine*, pages 1–10, 2023. 4, 7, 8, 10, 11

[22] Daniel P Huttenlocher, Gregory A. Klanderman, and William J Rucklidge. Comparing images using the hausdorff distance. *IEEE Transactions on pattern analysis and machine intelligence*, 15(9):850–863, 1993. 7

[23] Fabian Hörst, Moritz Rempe, Lukas Heine, Constantin Seibold, Julius Keyl, Giulia Baldini, Selma Ugurel, Jens Siveke, Barbara Grünwald, Jan Egger, and Jens Kleesiek. Cellvit: Vision transformers for precise cell segmentation and classification. *Medical Image Analysis*, 94:103143, 2024. 5, 6, 10

[24] 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. 3

[25] Talha Ilyas, Zubaer Ibna Mannan, Abbas Khan, Sami Azam, Hyongsuk Kim, and Friso De Boer. Tsfd-net: Tissue specific feature distillation network for nuclei segmentation and classification. *Neural Networks*, 151:1–15, 2022. 5

[26] Nicholas Konz, Yuwen Chen, Haoyu Dong, and Maciej A. Mazurowski. Anatomically-controllable medical image generation with segmentation-guided diffusion models. In *Proceedings of the Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, page pending. Springer Nature Switzerland, 2024. Accepted. 6

[27] Adrian B. Levine, Jason Peng, David Farnell, Mitchell Nursey, Yiping Wang, Julia R. Naso, Hezhen Ren, Hossein Farahani, Colin Chen, Derek Chiu, Aline Talhouk, Brandon Sheffield, Maziar Riazy, Philip P. Ip, Carlos Parra-Herran, Anne Mills, Naveena Singh, Basile Tessier-Cloutier, Taylor Salisbury, Jonathan Lee, Tim Salcudean, Steven J. M. Jones, David G. Huntsman, C. Blake Gilks, Stephen Yip, and Ali Bashashati. Synthesis of diagnostic quality cancer pathology images by generative adversarial networks. *The Journal of Pathology*, 252(2):178–188, 2020. 2

[28] Seonghui Min, Hyun-Jic Oh, and Won-Ki Jeong. Co-synthesis of histopathology nuclei image-label pairs using a context-conditioned joint diffusion model. In *Computer Vision – ECCV 2024*, pages 146–162, Cham, 2025. Springer Nature Switzerland. 3, 10

[29] 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 *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1999–2008, 2023. 2

[30] G. Müller-Franzes, J.M. Niehues, F. Khader, et al. A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis. *Scientific Reports*, 13:12098, 2023. 2

[31] Hyun-Jic Oh and Won-Ki Jeong. Diffmix: Diffusion model-based data synthesis for nuclei segmentation and classification in imbalanced pathology image datasets. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023: 26th International Conference, Vancouver, BC, Canada, October 8–12, 2023, Proceedings, Part III*, page 337–345, Berlin, Heidelberg, 2023. Springer-Verlag. 2, 3, 4, 5, 6, 7, 8, 10, 12

[32] Hyun-Jic Oh and Won-Ki Jeong. Controllable and efficient multi-class pathology nuclei data augmentation using text-conditioned diffusion models. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 36–46, Cham, 2024. Springer Nature Switzerland. 2, 3, 4

[33] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 6

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

[35] Adalberto Claudio Quiros, Roderick Murray-Smith, and Ke Yuan. Pathologygan: Learning deep representations of cancer tissue. In *Proceedings of the Third Conference on Medical Imaging with Deep Learning*, pages 669–695. PMLR, 2020. 2

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. 6, 8, 10, 11

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 2, 3, 4, 5, 10, 11

[38] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv:2010.02502*, 2020. 5

[39] Chetan L. Srinidhi, Ozan Ciga, and Anne L. Martel. Deep neural network models for computational histopathology: A survey. *Medical Image Analysis*, 67:101813, 2021. 1

[40] Yuxuan Sun, Chenglu Zhu, Sunyi Zheng, Kai Zhang, Lin Sun, Zhongyi Shui, Yunlong Zhang, Honglin Li, and Lin Yang. Pathasst: A generative foundation ai assistant towards artificial general intelligence of pathology. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38:5034–5042, 2024. 5, 6, 8, 10, 11, 13

[41] Luís A. Vale-Silva and Karl Rohr. Long-term cancer survival prediction using multimodal deep learning. *Scientific Reports*, 11(1):13505, 2021. Epub 2021 Jun 29. 1

[42] Gregory Verghese, Jochen K. Lennerz, Danny Ruta, Wen Ng, Selvam Thavaraj, Kalliopi P. Siziopikou, Threnesan Naidoo, Swapnil Rane, Roberto Salgado, Sarah E. Pinder, and Anita Grigoriadis. Computational pathology in cancer diagnosis, prognosis, and prediction - present day and prospects. *The*

Journal of Pathology, 260(5):551–563, 2023. Epub 2023 Aug 14. 2

[43] Ruchika Verma, Neeraj Kumar, Abhijeet Patil, Nikhil Cherian Kurian, Swapnil Rane, and Amit Sethi. Multi-organ nuclei segmentation and classification challenge 2020. *IEEE transactions on medical imaging*, 39(1380-1391):8, 2020. 5, 6, 8, 10, 11, 12

[44] Weilun Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Dong Chen, Lu Yuan, and Houqiang Li. Semantic image synthesis via diffusion models, 2022. 3, 4, 5, 6, 7, 8, 10, 12

[45] Srikar Yellapragada, Alexandros Graikos, Prateek Prasanna, Tahsin Kurc, Joel Saltz, and Dimitris Samaras. Pathldm: Text conditioned latent diffusion model for histopathology. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5182–5191, 2024. 2, 4, 10, 11

[46] Xinyi Yu, Guanbin Li, Wei Lou, Siqi Liu, Xiang Wan, Yan Chen, and Haofeng Li. Diffusion-based data augmentation for nuclei image segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 592–602, Cham, 2023. Springer Nature Switzerland. 3

[47] Lvmin Zhang, Maneesh K. Wu, Weiyang Zeng, Yuxin Zhang, Hussain Salman, and Vladlen Koltun. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 2, 4, 5, 6, 7, 8, 10, 12, 13