PathLDM: Text conditioned Latent Diffusion Model for Histopathology

Srikar Yellapragada * Alexandros Graikos * Prateek Prasanna Tahsin Kurc Joel Saltz
Dimitris Samaras

Stony Brook University

Abstract

To achieve high-quality results, diffusion models must be trained on large datasets. This can be notably prohibitive for models in specialized domains, such as computational pathology. Conditioning on labeled data is known to help in data-efficient model training. Therefore, histopathology reports, which are rich in valuable clinical information, are an ideal choice as guidance for a histopathology generative model. In this paper, we introduce PathLDM, the first text-conditioned Latent Diffusion Model tailored for generating high-quality histopathology images. Leveraging the rich contextual information provided by pathology text reports, our approach fuses image and textual data to enhance the generation process. By utilizing GPT's capabilities to distill and summarize complex text reports, we establish an effective conditioning mechanism. Through strategic conditioning and necessary architectural enhancements, we achieved a SoTA FID score of 7.64 for text-to-image generation on the TCGA-BRCA dataset, significantly outperforming the closest text-conditioned competitor with FID 30.1. [1]

1. Introduction

Diffusion models have achieved impressive performance across diverse applications such as image generation, text-to-video synthesis, and audio generation [13,14,22,42]. Recently, Latent Diffusion Models (LDM) [37] have paved the way for higher-resolution text-conditioned image generation by employing a two-step process involving image encoding and diffusion within the latent space.

Unsurprisingly, diffusion models in histopathology image generation are an active area of research. Existing works such as [28,29,48] are constrained by limited dataset size and conditioning capabilities. Our model performs a critical step toward the building of a *foundational model*, which is to accommodate multiple modalities, including images, text, class labels, etc.

Conditioning on labeled data improves the diffusion model's generation quality by enabling data-efficient model training. As shown in [31], conditioning can be particularly effective in refining diffusion models, especially when dealing with complex data distributions. Existing histopathology diffusion models condition on class labels, devoid of textual input. Due to the recent advances in Vision Language models [16, 27, 35], embeddings produced by VLMs are often more expressive than manually picked labels.

We introduce *PathLDM*, the first Latent Diffusion Model for text-conditioned histopathology image generation. Our novel approach hinges on employing pathology text reports as the conditioning signal. Histopathology data is multimodal, with pathology reports often accompanying whole slide images. These reports often include detailed descriptions of cell types, disease classification, and other domain-specific insights pathologists provide. They capture critical contextual information that is complementary to the diffusion model's generation ability.

We consider the extensively studied Cancer Genome Atlas Breast Invasive Carcinoma (TCGA-BRCA) dataset [1, 24]. Each entry within this dataset includes Whole Slide Images (WSI) and accompanying text reports provided by pathologists. Given that these reports can be lengthy (with a median length of 2800 characters) and unstructured, we adopt GPT's [32] capability to distill them into coherent summaries succinctly. These summaries serve as the guiding signal for PathLDM, effectively combining image and textual information. PathLDM produces visually coherent histopathology images aligned with the semantic details embedded within the text summaries through this synthesis.

Starting from the architecture of a Stable Diffusion [37] model, we propose a set of architectural modifications and develop PathLDM. Our technical modifications cover all three components of the LDM: the text encoder, the U-Net [38] denoiser, and the Variational Autoencoder (VAE).

For the conditioning to work, we use a cyclical positional embedding in the text encoder. We pick a VAE that preserves the intricate details, such as cell structures and spa-

*Equal contribution. Correspondence to srikary@cs.stonybrook.edu
[1]Code is available at https://github.com/cvlab-stonybrook/PathLDM

tial layout of primitives, found in pathology images. We choose the appropriate U-Net, which acts as the right initialization for our VAE. We demonstrate that our effective conditioning, coupled with the architectural refinements, enable PathLDM to achieve a state-of-the-art Fréchet Inception Distance (FID) score of **7.64** for text-to-image generation on the TCGA-BRCA dataset.

Comparing previous pathology diffusion models with Stable Diffusion, we observed that fine-tuned Stable Diffusion outperforms others. This can be attributed to Stable Diffusion being trained on a larger dataset. By conditioning with the appropriate pathology report text and the necessary architectural changes, we are the first to show that a pathology diffusion model can outperform fine-tuned Stable Diffusion.

Our contributions are as follows:

- We develop the first text-conditioned Latent diffusion model for histopathology image generation.
- We leverage GPT's capabilities to standardize and summarize pathology text reports for effective conditioning.
- We propose a framework that leverages conditioning with the appropriate architectural elements (cyclical positional embedding, VAE configuration, U-Net initialization) to improve the SoTA FID score dramatically.
- We are the first to demonstrate that a bespoke Pathology diffusion model can perform significantly better than Stable Diffusion.

2. Related Work

The seminal work of diffusion models for image generation [14] achieved comparable quality to ProgressiveGANs [17] in unconditional CIFAR-10 [23] synthesis. Building on this, [31] advanced DDPMs through class conditioning, learned variance and cosine noise scheduling, demonstrating competitive log-likelihoods in ImageNet [39] generation. Similarly, [8] improved the U-Net denoiser architecture with Adaptive Group Normalization, attention heads, and gradient-based sampling guidance, achieving SoTA results in ImageNet generation, rivaling BigGAN-deep [4].

Classifier-free diffusion guidance [15] combines unconditional and conditional score estimates to achieve performance akin to Classifier guidance without training an auxiliary classifier. The acceleration of diffusion sampling by 10 to 50 times is proposed in [43] through the Denoising Diffusion Implicit Models (DDIM). Latent Diffusion Models, introduced in [37], bifurcate the diffusion model training into two phases - an autoencoder for low-dimensional representation and a latent space diffusion model, thereby enhancing training efficiency. This approach surpasses previous methods in class conditional ImageNet generation and text conditional MS-COCO [26] synthesis [30]. Our paper is the first to develop a text-conditioned Latent Diffusion Model for histopathology images.

The use of pretrained diffusion models for synthetic data creation has gained attention recently. In [3], the authors employ a text conditional model on a large scale to create synthetic data for ImageNet. Similarly, [10] leverage features extracted from an unconditional diffusion model's U-Net to provide guidance and enhance synthetic data quality. Both adopt the Classification Accuracy Score [36] as a crucial metric to evaluate the quality of synthetic data.

The inception of CLIP: Contrastive Language Image Pretraining [35] marked a significant advancement in Computer Vision pre-training. Their approach involved jointly training image and text encoders to develop a multi-modal embedding space. For text conditional diffusion models such as Stable Diffusion [37], CLIP models were employed to encode text captions.

There have been recent advancements in Vision Language Models (VLMs) in the field of pathology. For instance, MI-Zero [27] leverages a dataset comprising over 33k image-text pairs to train their VLM. Similarly, PLIP [16] gathers pathology data from Twitter to train a CLIP model. The authors of BioMedCLIP [49] introduce domain-specific adaptations tailored to biomedical VLP to outperform prior VLP approaches. However, none of these methods have utilized the pathology image-text pairs to train diffusion models.

The application of diffusion models in histopathology remains limited. The authors of [28] train a pixel-level diffusion model for histopathology images, targeting 33,777 patches from low-grade gliomas within the TCGA dataset [11]. Similarly, [48] explore a Diffusion Autoencoder [34] trained on the NCT-CRC-HE [19] and PCam [45] datasets. Medfusion, introduced in [29], is a Latent diffusion trained on 19,558 colorectal cancer images from CRCDX [18].

Both [28] and [29] focus on training class conditional models. Meanwhile, [48] opts for separate unconditional models per class, which is less efficient than a shared class conditional model. It's worth noting that these approaches employ relatively small datasets and lack a text conditioning mechanism. Our PathLDM cures both of these deficiencies.

3. Methodology

In this section, we describe our approach to conditioning and the necessary modifications applied to Latent diffusion models for the histopathology setting.

3.1. GPT Summarization

We leverage the TCGA-BRCA dataset for our model training. Each entry within this dataset includes a Whole

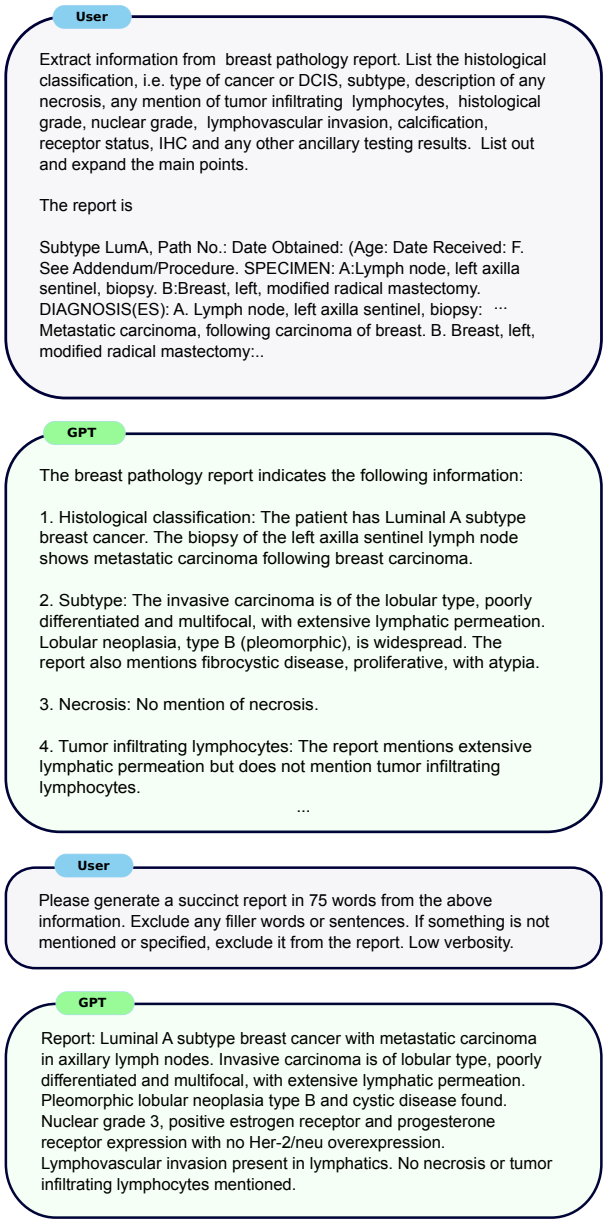

Figure 1. Example of our GPT-3.5 usage where we employ the chain-of-thought prompting technique to generate summaries. This approach leverages GPT's capability to progressively develop a coherent line of reasoning, yielding a more refined and informative summary.

Slide Image and an accompanying text report provided by pathologists, which we utilize as text conditioning. Accessible via cBioPortal [5, 9], these reports are primarily available in PDF format. Performing the complex process of Optical Character Recognition of PDFs to extract text is a non-trivial and challenging task. The authors of [20] tackled this challenge, extracting text for 9523 out of 11,000 TCGA text reports using OCR methods, further refined through manual validation. Their approach yielded a collection of extracted reports for 1034 BRCA WSIs, which we utilize.

The reports, on average, comprise around 2800 characters, roughly equivalent to 700 Byte-Pair Encoding tokens. These reports exhibit unstructured content due to the distinct writing style of each pathologist. SoTA text-to-image diffusion models, including Stable Diffusion, utilize a Transformer to encode the text captions, obtaining text embeddings. These embeddings are then cross-attended with image embeddings in the U-Net denoiser. Given that the cross-attention mechanism's computation complexity grows quadratically with sequence length, the need for concise summaries arises. Thus, we propose summarizing the reports and extracting crucial information using OpenAI's GPT model to ensure optimal processing within the model's limitations. It is worth noting that alternative LLMs, such as LLAMA [44], can also be considered for this purpose.

We use the OpenAI GPT API to summarize the reports. Specifically, we opt for the `gpt-3.5-turbo` variant, built upon GPT-3.5 and designed for optimized chat interactions. Zero-shot prompting is known to yield less consistent results than *chain of thought prompting*. As described in [47], this prompting technique generates a chain of thought through a series of intermediate reasoning steps and significantly improves the ability of Large Language models to perform complex reasoning.

Following this idea, we also observed that breaking down the summarization process into multiple sequential calls improves summarization quality (as verified by our pathologists). Our methodology involves initially requesting GPT to outline the main points within the report. We then prompt the model to generate a concise summary based on the rehashed key information. This approach leverages GPT's capability to progressively develop a coherent line of reasoning, yielding a more refined and informative summary. We randomly selected 5% of the generated summaries and had them validated by our board-certified pathologists.

In our second user query, we prompt GPT to generate a report using 75 words, approximately 150 tokens. Due to the nature of GPT, this does not guarantee that the response will indeed remain within the specified limit. To address this, we also impose a constraint (using the *max_tokens* parameter) on the final response generated by GPT, ensuring that it does not exceed 154 tokens. An example of an interaction that generates such a summary is presented in Fig. 1.

3.2. PathLDM

The core of our methodology revolves around the development of PathLDM, a Latent Diffusion Model tailored for histopathology image generation with text conditioning. PathLDM's architecture consists of the Variational Autoencoder (VAE), the U-Net denoiser and the text encoder. We provide an overview of our method in Figure 2.

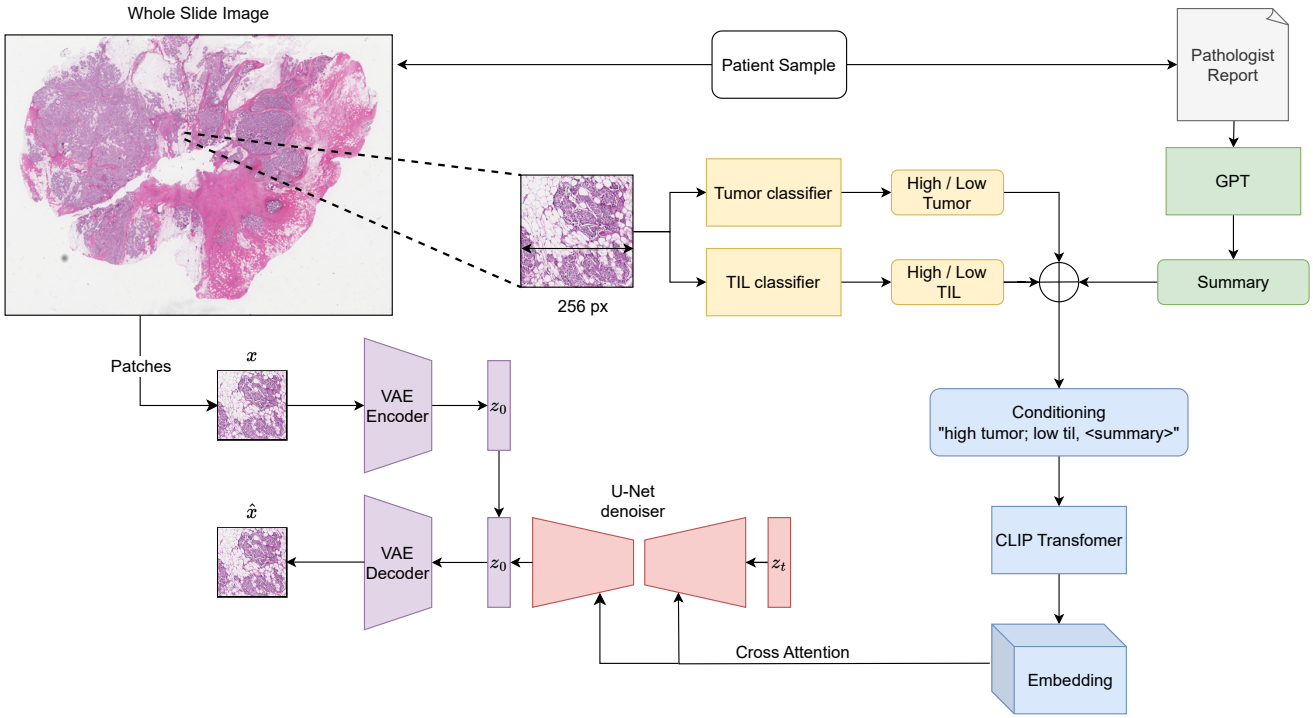

Figure 2. Overview of PathLDM. We start with a WSI and an accompanying Pathology report. We crop the WSI into 256 × 256 patches at 10x resolution. Leveraging GPT, we summarize the pathology report. For each patch, we compute Tumor and TIL probabilities, fuse them with the slide-level summary, and condition the LDM with CLIP embeddings of the fused summary. The VAE and text encoder remain frozen, and only the U-Net is trained.

3.2.1 Variational Autoencoder

The VAE learns a latent representation from the input images. A significant criterion for evaluating the VAE's performance is its ability to reconstruct the original image through the decoder. Accurate reconstructions are crucial, as the subsequent diffusion process operates on these latent vectors and can, therefore, be constricted by the VAE's capabilities.

Reconstruction quality becomes paramount when dealing with histopathology images that contain fine details such as cells. The authors of Medfusion [29] observed that using off-the-shelf Stable Diffusion's VAE can result in imperfect reconstructions. They modified the VAE architecture to overcome this issue by adjusting the number of channels and Convolution layer parameters. This refinement led to an increase in the Structural Similarity Index (SSIM) [46] of the reconstructed images.

However, both Stable Diffusion and Medfusion's VAE employ a downsampling factor of 8, effectively compressing 256 × 256 images down to a 32 × 32 × 4 (or 8 channels for Medfusion) latent. Opting for an overly small downsampling factor is undesirable, as it leads to a considerable increase in memory consumption and substantially hampers the training speed of the LDM. Thus, finding the right balance in compression levels is crucial for accurate reconstructions.

In PathLDM, we instead choose a 3-channel latent VAE with a downsampling factor of 4. In Section 4.3, we delve into the impact of the downsampling factor on VAE's reconstruction quality when fine-tuned on Pathology images. We experimentally validate the advantages of our latent size choice by showing how we can maintain the fine structures found in large images.

3.2.2 U-Net Conditioning

Leveraging GPT, we procure a single summary for each WSI. We then incorporate patch-level statistics corresponding to Tumor and Tumor-Infiltrating Lymphocyte (TIL) data to enrich these summaries. TILs have emerged as a promising biomarker in solid tumors, as a prognostic marker in triple-negative (TNBC) and HER2-positive breast cancer [2, 7, 40]. We use the trained models from [1] and [24] to obtain a Tumor and TIL probability value for each patch. These probability values were integrated into the text captions by first transforming them into an ordinal scale of Low - High and prepending them to the text description of each

patch. This process created captions like "High tumor; low til; {GPT summary}", effectively combining quantitative details with the summarization outcomes.

Our approach effectively fuses conditioning information derived from three distinct sources. By integrating the GPT summary, which encapsulates Whole-slide level conditioning, with Tumor and TIL statistics that represent patch-level information, we create an advantageous synthesis of global and local details. In contrast, existing Pathology diffusion models [28, 29, 48] exclusively rely on class conditioning mechanisms, devoid of textual input. We delve into the impact of our conditioning mechanism in Section 4.5.

3.2.3 Encoding large prompts

We use a CLIP [35] text encoder to process the text summaries obtained through GPT fused with Tumor and TIL presence. Its role is to transform these summaries into embeddings, compact yet rich textual context representations. Existing CLIP [16, 35] models are limited by a maximum context window length of 77 tokens. In our approach, we accommodate GPT summaries with a length of up to 154 tokens. This choice is deliberate, as it is twice the context length of the CLIP encoder. Opting for a length of 77 tokens would have resulted in an overly concise summary lacking in informative context.

To embed the full token sequence, we individually embed each of the halves and concatenate. This is equivalent to using a cyclical positional embedding that is repeated every 77 tokens, which we found to be sufficient for embedding the longer text sequences. We apply a causal mask and direct the full embedding sequence through the trained transformer encoder for processing. As an ablation, we replace the standard OpenAI CLIP [35] encoder with the more specialized PLIP encoder, discussed further in Section 4.6.

4. Experiments

In this section, we delve into the experimentation conducted to validate the effectiveness of our proposed PathLDM. We introduce architectural improvements to the Variational Autoencoder, U-Net denoiser, and text encoder components and explain how they affect the Fréchet Inception Distance (FID) metric.

We train all our diffusion models on 3 NVIDIA RTX 8000 GPUs with a batch size of 48 per GPU. We use the training code and checkpoints provided by [37], implemented in Pytorch [33]. Our training employs a learning rate of 2×10^{-5} with 10,000 warmup steps. For FID computation, we use the implementation from [41].

We adopt DDIM sampling [43] with 50 steps and incorporate classifier-free guidance [15] at a scale of 1.75 to generate synthetic images. We randomly sample classes for the class conditional models, and for the text conditional models, we randomly select summaries from the training set for conditioning.

4.1. Dataset

We selected 1136 Whole slide images from the TCGA-BRCA dataset and partitioned them into 977 for training and 199 for testing. Using the code from [25], we extract 256×256 patches at 10x magnification, culminating in 3.2 million training patches. Leveraging GPT, we procure a single summary for each WSI.

We use the trained models from [1] and [24] to obtain a Tumor and TIL probability value for each patch and incorporate them to enrich these summaries. This process created captions like "High tumor; low til; {GPT summary}", effectively combining quantitative details with the summarization outcomes.

4.2. Baseline

To assess the impact of different architectural changes on the FID, we begin with a class conditional LDM baseline. We exclusively utilize patch-level Tumor and TIL probabilities in this baseline without incorporating GPT's text summary. We first convert the probability values into an ordinal scale using a threshold of 0.5, denoted as "Low" and "High." Subsequently, each patch is assigned a class label from a range of values: 0 representing "Low Tumor + Low TIL", 1 for "Low Tumor + High TIL", 2 for "High Tumor + Low TIL", and 3 for "High Tumor + High TIL."

4.3. VAE

To quantify the influence of VAE on the diffusion model's generation quality, we train two class-conditional LDMs, leveraging patch-level Tumor and TIL data. One model uses a VAE with a downsampling factor of 8, denoted as "vq-f8", while the other employs a VAE with a downsampling factor of 4, denoted as "vq-f4". As exhibited in Table 1, FID shows noteworthy enhancement, improving from 48.14 to 39.72. This improvement underlines the substantial impact of VAE selection on the diffusion model's generative performance.

From the pretrained VAEs provided by [37], we start with VAE VQ-f4 with a downsampling factor of 4. We then fine-tune it on our image patches. In Table 2, we highlight the significance of the downsampling factor or latent vector dimension in the quality of VAE reconstructions. Our choice of VAE yielded a significant boost in the SSIM of the reconstructed images, suggesting that overly aggressive compression may hinder the retrieval of intricate details, particularly in images featuring small cellular structures.

Our observations underscore the vital role of VAE selection, even in other large-image settings such as satellite imagery. The Deepglobe Landcover dataset [6] comprises

VAE	f	U-net training	Conditioning network	Conditioning type	Conditioning modality	FID ↓
vq-f8	8	From scratch	Class embedder	Tumor + TIL	Class label (4 classes)	48.14
vq-f4	4	From scratch	Class embedder	Tumor + TIL	Class label (4 classes)	39.72
vq-f4	4	Finetune	Class embedder	Tumor + TIL	Class label (4 classes)	29.45
vq-f4	4	Finetune	OpenAI CLIP	Report + Tumor + TIL	Text caption (154 tokens)	10.64
vq-f4	4	Finetune	PLIP	Report + Tumor + TIL	Text caption (154 tokens)	**7.64**

Table 1. Effect of various components of our pipeline on FID. The use of text conditioning leads to the biggest reduction in FID. f: VAE's downsampling factor.

VAE	Latent size	f	SSIM ↑	MSE ↓
Stable Diffusion	32x32x4	8	0.874	25.795
Medfusion	32x32x8	8	0.891	21.401
Ours	64x64x3	4	**0.961**	**11.503**

Table 2. Comparison of reconstruction quality of VAE used in Stable Diffusion, medfusion and PathLDM, evaluated on TCGA-BRCA. The input image size is 256×256×3. A smaller downsampling factor (f) significantly boosts the SSIM.

VAE	Image Resolution	f	Latent size	SSIM ↑	MSE ↓
VQ-f4	256x256	4	64x64x3	0.958	2.359
	512x512		128x128x3	0.957	2.246
VQ-f8	256x256	8	32x32x4	0.845	6.235
	512x512		64x64x4	0.846	6.183

Table 3. Effect of latent size on the reconstruction quality of Satellite Images from Deepglobe [6] dataset. f: downsampling factor. The choice of VAE is significant even in diverse scenarios containing small objects.

803 RGB satellite images, each measuring 2448x2448 pixels. We partition these images into 256x256 and 512x512 crops and reconstruct them using pre-trained VAEs sourced from [37]. The results presented in Table 3 provide compelling evidence that opting for a reduced downsampling factor substantially improves the SSIM of reconstructed images. This finding further highlights the significance of VAE choice in diverse image domains, including scenarios involving smaller-sized objects such as satellite imagery.

4.4. U-Net

Moving forward, our next focus is the U-Net denoiser. To gauge the advantage of fine-tuning, we train two class conditional LDMs. In the first model, we initialize the U-Net denoiser weights from scratch. Second, we start with the ImageNet [39] class conditioned U-Net denoiser weights from [37]. The outcome presented in Table 1 shows that employing a fine-tuned U-Net enhances the FID from 39.72 to 29.45. Intriguingly, even though the ImageNet weights are tailored to Natural images - distinct from the complexity of histopathology images - the learned denoising mechanism within the U-Net contributes significantly to the efficacy of our PathLDM training.

4.5. Text conditioning

To measure the concrete effects of our VAE and U-Net changes, we initially eshtablished a class conditional LDM baseline. Now, we aim to assess the benefit of introducing text summaries. Leveraging GPT, we obtain a summary for each Whole slide image (WSI). These summaries are combined with patch-level Tumor and TIL classes and employed to condition the diffusion model.

The results, as outlined in Table 1, are impressive: the combination of text reports and patch-level statistics significantly reduces the FID score from 29.45 to 10.64. This improvement outperforms the gains achieved solely through architectural enhancements. Incorporating context from text summaries introduces a powerful interplay between visual and textual data, substantially enhancing model performance.

4.6. Text encoder

Stable Diffusion uses the OpenAI CLIP [35] encoder, trained on Natural image caption pairs. As an ablation experiment, we introduce PLIP [16], a CLIP model trained on pairs of pathology images and their corresponding text. As shown in Table 1, adopting the PLIP encoder improved the FID from 10.64 to **7.64**. This improvement is attributed to the PLIP model's more adept grasp of the nuances present in pathology-specific text summaries.

4.7. Comparison with other methods

The cumulative architectural enhancements detailed in the preceding subsections result in the creation of our most potent Latent Diffusion Model - PathLDM. We compare our method against three alternative methods - Moghadam et al. [28], Medfusion [29] and Stable Diffusion [37].

For Moghadam et al. and Medfusion, both class conditional diffusion models, we train them with Tumor and TIL classes. Similarly, we train Stable Diffusion using text summaries, akin to our approach for PathLDM. We initialize the U-Net denoiser of Stable Diffusion and PathLDM

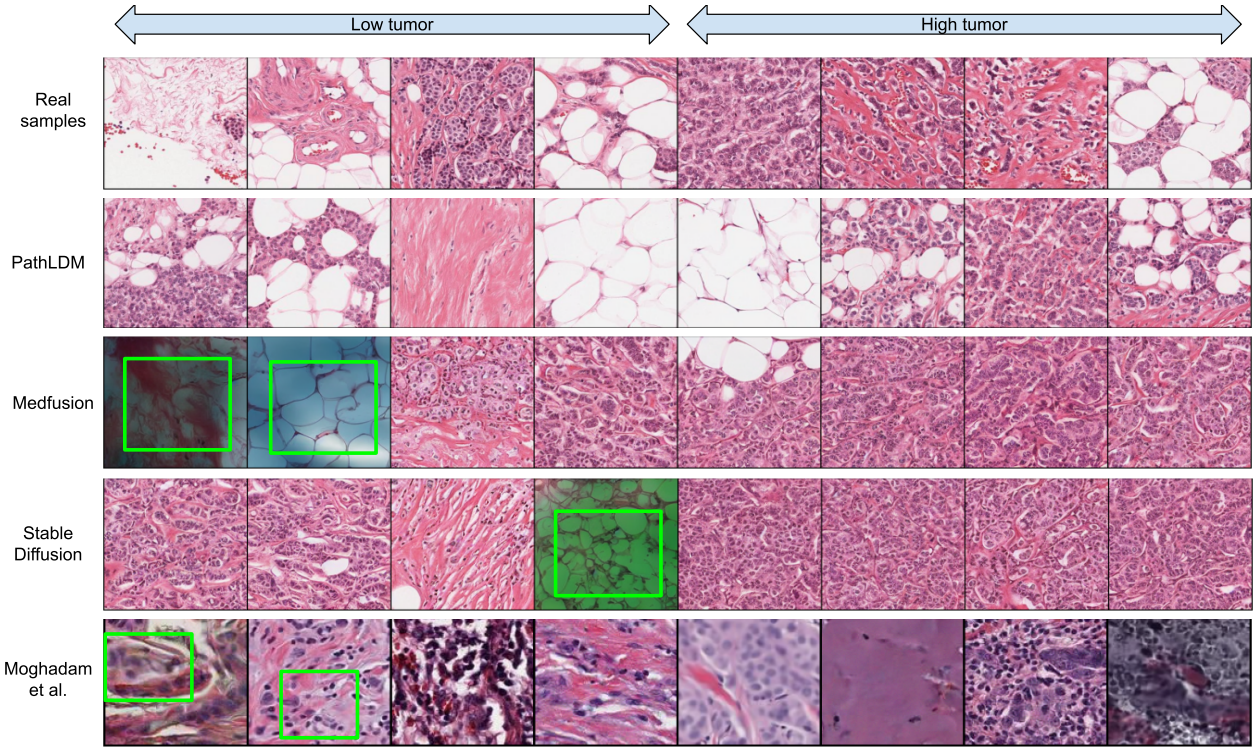

Figure 3. We choose a single text report and produce synthetic samples using Medfusion [29], Stable Diffusion [37], and PathLDM. Samples generated by Medfusion and Stable Diffusion show artifacts (indicated by green boxes) that are not present in our outputs. Moghadam et al. [28] produces images in lower resolution (128×128) and exhibits artifacts and blurriness (green boxes in the last row) in the output images. The first row contains the real samples from the corresponding WSI.

Method	Conditioning type	FID ↓
Moghadam et.al [28]	Tumor + TIL	105.81
Medfusion [29]	Tumor + TIL	46.49
	Report + Tumor + TIL	39.56
Stable Diffusion [37]	Report + Tumor + TIL	30.1
PathLDM	Tumor + TIL	29.45
	Report + Tumor + TIL	**7.64**

Table 4. Comparison of FID using 10k synthetic images. PathLDM outperforms other methods by a huge margin. Training Medfusion with text-conditioning improved the FID, highlighting the efficacy of our text summaries.

using pretrained weights from [37]. However, Medfusion employs an 8-channel VAE, and no existing pretrained U-Net for Stable Diffusion has done more than four channels, which makes them unsuitable for initialization. Hence, we train it from scratch.

The results, as shown in Table 4, demonstrate the superiority of our text conditional PathLDM over other methods, establishing a significant FID gap. Furthermore, we also conducted text-conditional training for the Medfusion architecture, leading to an FID improvement from 46.49 to 39.56, highlighting the efficacy of our text summaries. Notably, even our class conditional model, trained without text reports, outperforms the class conditional variants of both Moghadam et al. and Medfusion, owing to the potency of the architectural enhancements we introduced.

Pixel-level diffusion models are computationally demanding, and no straightforward method exists to condition them on text captions. Recent works like GLIDE [30] and Imagen [13] are typically constrained to smaller image sizes (64×64) and rely on upsampling models to achieve high-resolution image synthesis. Since Moghadam et al. utilizes a pixel-level diffusion model, we haven't trained a text-conditional variant.

For visualization, we select a text report and generate synthetic samples by employing Medfusion, Stable Diffusion, and PathLDM. Four samples each, categorized as "low tumor" and "high tumor," are generated for evaluation. The figures in Figure 3 highlight the diversity and authenticity of the samples produced by our approach. Notably, our method generates samples that closely adhere to the distribution of actual samples. On the other hand, samples generated by Medfusion and Stable Diffusion show artifacts, such as blue stains, that are absent in our outputs.

Data type	Data source	Accuracy (%)
Real	TCGA-BRCA	90.31
Only Synthetic	Medfusion	79.12
Only Synthetic	PathLDM	**83.32**
Real + Synthetic	PathLDM	**90.74**

Table 5. Classification Accuracy scores of a ResNet-34 on synthetic data. Higher CAS proves the effectiveness of PathLDM-generated synthetic samples in capturing significant features.

PathLDM Variant	FID
Baseline - True summary	7.64
Random summary	22.16
5 summaries per WSI	7.34

Table 6. Ablation to highlight the benefit of text summaries.

4.8. Classification Accuracy Score

In addition to the widely used FID for evaluating generative models, an alternative evaluation metric is the Classification Accuracy Score (CAS) [36]. To compute CAS, a classifier is trained on synthetic data and evaluated on real validation data. To train PathLDM, we use a dataset of 3 million patches, each associated with a Tumor label of either "Low" or "High." We create synthetic datasets using PathLDM and Medfusion, generating 3 million patches. We assign the Tumor label used in generating each patch as its ground truth.

We then train ImageNet pretrained ResNet-34 [12] classifiers on these synthetic datasets for 40 epochs with a batch size 256 and the Adam optimizer [21]. We use a learning rate of 10^{-3} for the first 20 epochs, then a reduction to $\frac{1}{10}$th at epochs 20 and 30. As presented in Table 5, the classifier trained on our synthetic data achieves a validation accuracy of **83.32%**, significantly higher than a classifier trained on Medfusion's data (79.12%). When the classifier is trained using a mix of real and synthetic data, there is a modest improvement in accuracy, increasing from 90.31% to **90.74%**. This marginal enhancement suggests that the task of tumor classification is inherently less complex.

5. Ablation studies

5.1. Sampling parameters

We adopt DDIM sampling with 50 steps and integrate classifier-free guidance. We conduct a parameter sweep across guidance values ranging from 0.1 to 3, resulting in the optimal FID outcome at 1.75. This value is then employed consistently across all our experimental scenarios.

5.2. Benefit of text reports

We perform two ablation experiments to underscore the significance of text report conditioning. In the first ablation, we train an LDM where, for each patch, we substitute its Whole Slide Image (WSI) summary with a randomly chosen summary from a different WSI. In the second ablation, we extend the utility of GPT by generating five distinct summaries from each text report obtained through five separate GPT API queries. Each of these queries operates independently and adheres to the same token limitations. Subsequently, for each patch, we randomly select only one summary from the five provided by GPT.

The outcomes, as shown in Table 6, provide valuable insights - substituting a random summary leads to the deterioration of the diffusion model, increasing the FID from 7.64 to 22.16. This outcome demonstrates that using an inappropriate summary or the absence of one altogether worsens the generation quality. Conversely, deploying five summaries per WSI improved the FID from 7.64 to **7.34**. These ablations underline the crucial role that accurate text conditioning plays in enhancing the performance of our model.

6. Limitation

A limitation of our approach is the reliance on text prompts from the training set for generating synthetic images. We did not employ held-out reports for image generation, which could be a potential avenue for future research.

7. Conclusion

Our paper presents PathLDM, a groundbreaking Latent Diffusion Model designed for generating histopathology images conditioned on text. By effectively summarizing complex pathology text reports using GPT, we establish a robust conditioning mechanism that bridges the gap between textual and visual content. Our strategic conditioning, coupled with architectural enhancements, led to a SoTA FID score of 7.76 on the TCGA-BRCA dataset, significantly outperforming the closest competitor with FID 30.1. This study performs a critical step toward the building of a foundational model, which is accommodating multiple modalities, including images, text, class labels, etc. We anticipate that our work will encourage future exploration of domain-specific pathology diffusion models.

Acknowledgements

The reported research was partially supported by NCI awards 5U24CA215109, 1R21CA258493-01A1, NSF grants IIS-2123920, IIS-2212046 and Stony Brook Profund 2022 seed funding.

References

[1] Shahira Abousamra, Rajarsi Gupta, Le Hou, Rebecca Batiste, Tianhao Zhao, Anand Shankar, Arvind Rao, Chao Chen, Dimitris Samaras, Tahsin Kurc, et al. Deep learning-based mapping of tumor infiltrating lymphocytes in whole slide images of 23 types of cancer. *Frontiers in oncology*, 11:806603, 2022. 1, 4, 5

[2] Mohamed Amgad, Elisabeth Specht Stovgaard, Eva Balslev, Jeppe Thagaard, Weijie Chen, Sarah Dudgeon, Ashish Sharma, Jennifer K Kerner, Carsten Denkert, Yinyin Yuan, et al. Report on computational assessment of tumor infiltrating lymphocytes from the international immuno-oncology biomarker working group. *NPJ breast cancer*, 6(1):16, 2020. 4

[3] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*, 2023. 2

[4] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 2

[5] Ethan Cerami, Jianjiong Gao, Ugur Dogrusoz, Benjamin E Gross, Selcuk Onur Sumer, Bülent Arman Aksoy, Anders Jacobsen, Caitlin J Byrne, Michael L Heuer, Erik Larsson, et al. The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer discovery*, 2(5):401–404, 2012. 3

[6] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. DeepGlobe 2018: A challenge to parse the earth through satellite images. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, jun 2018. 5, 6

[7] Carsten Denkert, Sibylle Loibl, Aurelia Noske, Marc Roller, BM Muller, Martina Komor, Jan Budczies, Silvia Darb-Esfahani, Ralf Kronenwett, Claus Hanusch, et al. Tumor-associated lymphocytes as an independent predictor of response to neoadjuvant chemotherapy in breast cancer. *J Clin Oncol*, 28(1):105–113, 2010. 4

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

[9] Jianjiong Gao, Bülent Arman Aksoy, Ugur Dogrusoz, Gideon Dresdner, Benjamin Gross, S Onur Sumer, Yichao Sun, Anders Jacobsen, Rileen Sinha, Erik Larsson, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal. *Science signaling*, 6(269):pl1–pl1, 2013. 3

[10] Alexandros Graikos, Srikar Yellapragada, and Dimitris Samaras. Conditional generation from unconditional diffusion models using denoiser representations. *arXiv preprint arXiv:2306.01900*, 2023. 2

[11] Robert L Grossman, Allison P Heath, Vincent Ferretti, Harold E Varmus, Douglas R Lowy, Warren A Kibbe, and Louis M Staudt. Toward a shared vision for cancer genomic data. *New England Journal of Medicine*, 375(12):1109–1112, 2016. 2

[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. 8

[13] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 1, 7

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 2

[15] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2, 5

[16] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas Montine, and James Zou. Leveraging medical twitter to build a visual–language foundation model for pathology ai. *bioRxiv*, pages 2023–03, 2023. 1, 2, 5, 6

[17] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 2

[18] Jakob Nikolas Kather. Histological image tiles for tcga-crc-dx, color-normalized, sorted by msi status, train/test split, May 2020. 2

[19] Jakob Nikolas Kather, Niels Halama, and Alexander Marx. 100,000 histological images of human colorectal cancer and healthy tissue, Apr 2018. 2

[20] Jenna Kefeli and Nicholas Tatonetti. Benchmark pathology report text corpus with cancer type classification. *medRxiv*, pages 2023–08, 2023. 3

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 8

[22] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020. 1

[23] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 2

[24] Han Le, Rajarsi Gupta, Le Hou, Shahira Abousamra, Danielle Fassler, Luke Torre-Healy, Richard A Moffitt, Tahsin Kurc, Dimitris Samaras, Rebecca Batiste, et al. Utilizing automated breast cancer detection to identify spatial distributions of tumor-infiltrating lymphocytes in invasive breast cancer. *The American journal of pathology*, 190(7):1491–1504, 2020. 1, 4, 5

[25] 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. 5

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2

[27] Ming Y. Lu, Bowen Chen, Andrew Zhang, Drew F. K. Williamson, Richard J. Chen, Tong Ding, Long Phi Le, Yung-Sung Chuang, and Faisal Mahmood. Visual language pretrained multiple instance zero-shot transfer for histopathology images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19764–19775, June 2023. 1, 2

[28] 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. 1, 2, 5, 6, 7

[29] 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. 1, 2, 4, 5, 6, 7

[30] Alex 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. *arXiv preprint arXiv:2112.10741*, 2021. 2, 7

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

[32] OpenAI. OpenAI: Introducing ChatGPT, 2022. 1

[33] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 5

[34] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *CVPR*, pages 10609–10619. IEEE, 2022. 2

[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. 1, 2, 5, 6

[36] Suman Ravuri and Oriol Vinyals. Classification accuracy score for conditional generative models. *Advances in neural information processing systems*, 32, 2019. 2, 8

[37] 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, 5, 6, 7

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 1

[39] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 2, 6

[40] Peter Savas, Roberto Salgado, Carsten Denkert, Christos Sotiriou, Phillip K Darcy, Mark J Smyth, and Sherene Loi. Clinical relevance of host immunity in breast cancer: from tils to the clinic. *Nature reviews Clinical oncology*, 13(4):228–241, 2016. 4

[41] Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid, August 2020. Version 0.3.0. 5

[42] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 1

[43] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 5

[44] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 3

[45] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*, pages 210–218. Springer, 2018. 2

[46] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 4

[47] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022. 3

[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. 1, 2, 5

[49] 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. 2