ISSN-L 2664-9969; E-ISSN 2706-5448

UDC 004.8:004.932.72:616-091.8
DOI: 10.15587/2706-5448.2026.355663

Sergii Kuzmin,
Oleh Berezsky

DEVELOPMENT OF A PARAMETER-EFFICIENT METHOD FOR BIOMEDICAL IMAGE SYNTHESIS BY SUBSTITUTING TEXT CONDITIONING WITH PATHOLOGY FOUNDATION MODEL EMBEDDINGS IN LATENT DIFFUSION

The object of research is the process of synthesizing patches of histopathological images conditioned by embeddings of the pathology foundation model. One of the key problems is that existing approaches to diffusion synthesis either rely on text conditioning via CLIP encoders, which lack morphological understanding, or require full retraining of the generative base model, which requires significant computational resources.

The research used a parameter-efficient adaptation of the previously trained latent diffusion model using low-rank adaptation (LoRA) of the U-Net attention layers in combination with a training MLP projector that reflects the embeddings of the pathology foundation model UNI2-h in the conditioning space of the cross-attention mechanism. Ablation studies of 12 configurations were conducted varying the adapter rank, the number of conditioning tokens, and the projector architecture.

It is confirmed that embeddings of the pathology foundation model can effectively replace text conditioning for the synthesis of histopathology images in a parameter-efficient mode. The optimal configuration achieved FID 77.59 on the validation set and FID 84.17 on the test set when training only 5.53 million parameters, which is 0.64% of the parameters of the base model. This is due to the fact that the proposed method has a number of characteristic features, in particular: embeddings of the pathology foundation model provide morphologically richer conditioning than CLIP-based text representations, and low rank adaptation limits the trainable space to the conditioning pathway.

This provides the possibility of generating histopathology images without text annotations and without full retraining of the model using approximately 12 GB of video memory. Compared to the previous text-conditioned approach on the same dataset, which demonstrated class-wise FID values in the range of 113 to 138, the embedding conditioning method provides significantly higher generation quality while maintaining parameter efficiency.

Keywords: *latent diffusion models, pathology foundation models, histopathology image synthesis, medical image generation.*

Received: 05.01.2026
Received in revised form: 04.03.2026
Accepted: 25.03.2026
Published: 30.04.2026

How to cite
Kuzmin, S., Berezsky, O. (2026). Development of a parameter-efficient method for biomedical image synthesis by substituting text conditioning with pathology foundation model embeddings in latent diffusion. Technology Audit and Production Reserves, 2 (2 (88)), 66–75. https://doi.org/10.15587/2706-5448.2026.355663

1. Introduction

The cost of obtaining the high-quality, annotated image datasets required for automated histopathology analysis is a problem. Large, well-annotated datasets take a great deal of expertise and a lot of time to create: a single WSI is a gigapixel-scale image. Labeling such image at the pixel level will take hours of an expert's time per slide. For example, the CAMELYON benchmark is illustrative of this difficulty – 1,399 annotated WSIs totaling approximately 3 terabytes were produced in a collaborative multi-institutional effort that is rare outside of benchmark projects [1].

In addition to workforce and regulatory issues, there are other constraints limiting the availability of histopathology data. Several recent reviews have documented ongoing pathologist shortages and increasing complexity of diagnosis in major health care systems [2]. Privacy laws (HIPAA/GDPR) and institutional governance policies limit the ability to share patient-derived data [3]; thus, training datasets are fragmented across institutions and jurisdictions. Additionally, histopathology datasets suffer from extreme class imbalance, so that rare cancer subtypes appear very infrequently compared to common tissue types, which causes models to systematically favor the majority classes [4, 5]. This is why the use of synthetic data for generating additional examples of underrepresented classes and rare morphologies is considered a useful complementary approach to directly sharing patient data. However, the privacy risk of memorization needs to be evaluated when deploying synthetic data pipelines.

To generate realistic tissue images for data augmentation purposes, generative models satisfy this requirement. Medical image synthesis, including histopathology, has primarily employed GAN-based and class/mask conditioned models [6]. However, these models typically

require task-specific annotation and can be difficult to scale to diverse, high-resolution morphology. Partly it is caused by their tendency to experience unstable training and/or poor mode coverage compared to diffusion backbones [7, 8]. Recently, latent diffusion models (LDMs) [9] have emerged as a broadly adopted paradigm for generating high-fidelity images of biological tissues, achieving impressive performance while working in a compressed latent space. However, adapting LDMs to new specialized domains usually requires significant modification of the substantial portion of the U-Net architecture, which is computationally expensive. Standard stable diffusion is specifically designed for generating images conditioned on the output of a pre-trained text encoder [9, 10]. Typically, proxy captions are created through report-to-captions pipelines; however, there is a one-to-many relationship between the nuances of morphology and short textual descriptions. In particular, describing the distinctions between well-differentiated and poorly differentiated adenocarcinomas, in terms that pre-trained text encoders can reliably detect, is difficult. Features that contribute to diagnoses arise from subtle variations in cellular architecture and staining patterns that are difficult to describe concisely verbally. Thus, text conditioning offers relatively little flexibility in controlling fine-grain morphology and stain-dependent features in histopathology.

Self-supervised pathology foundation models present an alternative conditioning signal that is more directly related to tissue morphology. Self-supervised pathology models, such as UNI and UNI2-h [11], have been trained on large WSI collections using self-supervised objectives. These models produce rich, task-agnostic representations of morphology that contain diagnostically relevant information without requiring explicit labels. Since these representations of morphology exist in a compact, semantically meaningful format, they serve as a natural control signal for guiding the synthesis process and avoid the limitations associated with text.

Recent studies have explored the use of diffusion-based histopathology synthesis using several different conditioning strategies; however, each of these approaches has its own limitations. Text-conditioned methods, such as the one described in [12], utilize captions generated from reports or metadata; thus, they inherit the limitations of the text encoder. Representation-guided method in paper [13] demonstrates the feasibility of utilizing self-supervised representations to condition the synthesis process and fine-tune ImageNet pre-trained U-Net weights. However, such methods require significant computational resources and do not leverage parameter-efficient adaptation strategies. Hybrid method proposed in paper [14] utilizes parameter-efficient LoRA adaptation [15] to adapt the pre-trained Stable Diffusion; however, they retain the text encoder and therefore require text prompts. No previous study has been found that utilizes both parameter-efficient adaptation of a pre-trained latent diffusion backbone and pathology foundation model embeddings as the sole conditioning signal without relying on a text encoder.

This gap is addressed through focus on patch-level synthesis, and using pathology foundation embeddings computed from input tiles to condition generation. Morphology-preserving data augmentation may be enabled via generating many variations of images reflecting underrepresented classes or infrequently observed morphologies; each sample varies due to random sampling in the diffusion model. The dataset is divided so that there is no overlap between the sets used to train and evaluate the models, thereby preventing the potential of information leakage. A lightweight projector replaces the original text encoder, projecting these pathology foundation embeddings to create cross-attention conditioning tokens, and LoRA adapters are inserted into the U-Net's attention layers. Compared to previous methods, suggested approach requires only a very small number of backbone parameters to be updated, enabling the generation of pathology-conditioned images without text-based prompts. Unlike text-based prompts, the conditioning signals are generated automatically based on the tissue morphology using a pre-trained pathology encoder. This eliminates the requirement for manually authored captions or the need for human engineers to design prompts.

Systematic ablation studies have been performed on LoRA rank, the number of conditioning tokens, projector architectures and the learning rate ratio to quantify the trade-offs between compute and quality. Proposed method can be trained on a modern consumer GPU (e. g., with 16 GB of memory) depending on the resolution and batch size and therefore provides practical guidelines for deployment in limited-resource environments. As such, this method should be viewed not as a replacement for curated real-world datasets, but as complementary to them.

The application of generative adversarial networks (GANs) to histopathology image synthesis is a theme represented in a growing body of literature, indicating the success of GANs for data augmentation, with several authors observing increases in classification accuracy across various downstream classification tasks [6]. However, as mentioned earlier, one of the primary drawbacks of GANs is their tendency to unstable training and mode collapse when generating a wide range of tissue patterns. In paper [7], the authors reviewed many issues associated with training GANs in biomedical imaging and noted that mode collapse remains a significant problem, leading to fewer varied tissue patterns. This has resulted in researchers looking into other approaches to image synthesis including diffusion models.

In recent years, diffusion models have emerged as a dominant form of high-fidelity image synthesis. A popular formulation was presented in paper [16] in which it was demonstrated that denoising iterations could produce high-quality samples, comparable to GANs. More recently, in paper [8] the authors presented that diffusion models can either match or exceed GANs on standard benchmarks with FID scores of 2.97 on class-conditional ImageNet 128x128 and 4.59 on ImageNet 256 × 256. In order to address the high computational cost of performing diffusion in the pixel space, latent diffusion models (LDMs) that perform denoising in a compact latent space were developed [9].

Stable diffusion uses both a spatial autoencoder for compressing the input image and a U-Net-based denoiser. As outlined in the LDM framework [9], a general cross-attention conditioning mechanism was introduced that accommodates diverse types of conditioning signals. Many popular implementations of Stable Diffusion, such as those using CLIP-based text encoders [10], generate embeddings that are then integrated through cross-attention layers in the U-Net. However, CLIP-based text encoders may be poorly suited for histopathology because the semantic gap between visual morphology and natural language significantly limits the precision and fine-grained morphological distinctions achievable through conditioning.

Pathology foundation models offer an alternative conditioning paradigm for medical image synthesis. In paper [11], UNI, a self-supervised model pretrained on more than 100 million images from over 100,000 diagnostic whole-slide images stained with Hematoxylin and Eosin across 20 main tissue types, was introduced. The UNI model was shown to be the best-performing model to date for 34 clinical tasks; it was better than prior encoders and greatly exceeded the performance of a ResNet-50 model trained on ImageNet (for example, +26.4% average performance on 15 slide-level tasks and +18.8% average performance on 11 ROI-level tasks). Following the success of the UNI model, the UNI2-h model employed a ViT-H/14 architecture with 681M parameters and produced 1536-dimensional embeddings. These embedding representations were designed to be task agnostic and to represent structural elements of the tissue. As such, the UNI and UNI2-h models have the ability to provide very rich embeddings that may be useful as conditioning signals for many other machine learning applications. However, their utility as conditioning signals has not been systematically explored for conditioning pre-trained latent diffusion backbones via parameter-efficient adaptation.

ISSN-L 2664-9969; E-ISSN 2706-5448

Although some recent works have focused on developing diffusion models specific to histopathology generation, due to differences in how each study evaluates their results, in this study reported FID scores are used to illustrate trade-offs rather than to compare them. In paper [12], PathLDM, a latent diffusion model fully fine-tuned and conditioned on text for generating histopathology images, was presented. They reported a FID score of 7.64 on TCGA-BRCA at a resolution of 256 × 256. However, since PathLDM uses CLIP-based encoding for text conditioning, it necessitates summarization of lengthy pathology reports and can create an information bottleneck. In contrast, LRDM, conditioned on self-supervised embeddings (HIPT for BRCA and iBOT for CRC) was presented in [13] with reported patch-level FID of 6.98 on TCGA-BRCA. While LRDM used components of the LDM framework to train the model, training is computationally expensive; it requires the full fine-tuning of the U-Net backbone using six RTX 8000 GPUs and 15 million training patches. In [17], this line of research was further advanced with the introduction of ZoomLDM for multi-scale synthesis, reporting a FID of 6.77 at the 20x magnification level. Similarly to LRDM, full fine-tuning of the U-Net backbone is necessary for training this model.

The use of parameter-efficient fine-tuning is a way to reduce computation when adapting diffusion models for different applications. LoRA method was introduced in [15] where the authors froze pre-trained weights and added trainable low-rank matrices, resulting in performance that was equivalent to full fine-tuning while greatly reducing the number of trainable parameters (i. e., 10,000× on GPT-3 175B). Some recent modifications to the original LoRA method include SeLoRA [18], which proposes a dynamic rank expansion for medical image applications. They found that this resulted in a decrease of greater than 50% in FID compared to the same fixed rank LoRA used on small datasets (Montgomery County CXR), while producing comparable results to large-scale datasets (IU X-RAY).

In paper [14] it is showed how stable diffusion could be adapted for cytology synthesis using LoRA adaptation, which increased the accuracy of a downstream classifier from 27% to 78%. However, the authors still utilized the text-based conditioning mechanism of the stable diffusion text encoder, which requires either manually generated text prompts or automatically generated text prompts.

In a related line of work, the development of biomedical image datasets containing real cytological, histological, and immunohistochemical images was described in [19]. The application of generative intelligence tools for synthesizing biomedical images was further investigated in [20], addressing the practical aspects of artificial image generation for expanding biomedical datasets. In a complementary direction, an automated method for searching optimal convolutional neural network architectures for biomedical image classification was proposed in [21], which outperformed well-known networks such as VGG-16 and AlexNet on oncological imaging tasks. Furthermore, a method and cloud-based software tool for generating artificial databases of biomedical images using GANs was presented in [22].

Prior research described in [23] covers Stable Diffusion adaptation which utilizes class-label text prompts in order to generate images of colon histopathology; results included per-class FID scores ranging from 113 to 138. While effective for initial exploration, this text-based conditioning motivated the present investigation into pathology foundation model embeddings as an alternative signal source.

Despite these advances in parameter efficiency there is no evidence that these methods have ever been combined with pathology foundation model embeddings to create conditioned histopathology synthesis.

This review clearly identifies a methodological gap: embedding conditioned histopathology diffusion has been developed without utilizing parameter-efficient adaptation strategies of the type described above, while parameter-efficient adaptation strategies, like LoRA, continue to be dependent upon text encoders. These observations motivate a unified approach that utilizes LoRA adaptation and pathology foundation model embeddings as the single source of conditioning information.

The object of this research is the process of synthesizing histopathology image patches conditioned on pathology foundation model embeddings.

The aim of this research is to demonstrate that pathology foundation model embeddings can be used instead of text conditioning to generate synthetic histopathology images in parameter-efficient adaptations. This would allow deployment on consumer-grade hardware, and eliminate the need for textual supervision or manually writing prompts. To support this demonstration, systematic ablation studies across major architectural decisions provide empirical guidance for quality-compute trade-off decisions in resource-constrained environments.

In order to achieve the aim, the following objectives are set:

– to design an embedding-to-cross-attention projector architecture that, combined with low-rank adaptation (LoRA) of the U-Net attention layers, enables parameter-efficient fine-tuning of the baseline diffusion model and demonstrates the feasibility of pathology foundation model conditioning;

– to perform a systematic ablation study on major architecture decisions (LoRA rank, number of conditioning tokens, projector depth & hidden dimensionality, learning rate ratio between the adapter and projector);

– to evaluate the quality of generated images using the Fréchet inception distance (FID) and discover the optimal hyperparameters that best balance image quality and computational efficiency.

2. Materials and Methods

Unlike previous approaches, which use either the full fine-tune of a pre-trained backbone or text-based conditioning, this paper uses parameter-efficient adaptation of a pre-trained latent diffusion model to investigate whether self-supervised representations of a pre-trained backbone could be used as a conditioning signal when paired with a low-rank adapter.

The central hypothesis of this research is that pathology foundation model embedding will be able to substitute text conditioning for generating histopathology images with a learned projector and low rank adaptation; thus, allowing for the generation of morphology-aware images without the use of a text encoder in addition to reducing computational requirements. It is based on the following assumptions:

– UNI2-h embeddings capture diagnostically relevant morphological features, which will provide a richer conditioning signal than CLIP-based text embeddings for histopathology due to the fact that these representations were learned from large whole-slide image collections using self-supervised objectives instead of natural language supervision;

– low-rank adaptation of U-Net attention layers will be sufficient to convert a text-conditioned diffusion backbone to an embedding-conditioned generation model, without the need to perform full fine-tuning, since the pre-trained weights of the diffusion backbone already contain a general understanding of both image structure and denoising dynamics;

– systematic variation of architectural hyperparameters (LoRA rank, No. of conditioning tokens, projector depth & hidden dimension, learning rate ratio), will allow to evaluate quality-efficiency trade-offs and make informed decision regarding deployment based on differing levels of resource availability;

– fréchet inception distance (FID) computed between real and synthetic images will provide a meaningful proxy for measuring the quality of generated images, and enable quantitative comparisons among configurations.

However, a few simplifications were made in the research process, specifically:

– a fixed resolution of 512 × 512 pixels was used for all experiments, since this matches the native training resolution of stable diffusion 1.5, and eliminates the added complexity of multi-scale generation;

– both the pre-trained pathology encoder (UNI2-h) and the variational autoencoder (VAE) remained frozen throughout training; the only trainable components were the projector and the LoRA adapters, which limit the experimental space to the components that mediate the conditioning pathway;

– evaluation was restricted to a single dataset; the evaluation of downstream classification performance was deferred until future work, in order to focus this research on establishing the feasibility of the proposed conditioning methodology.

Experiments were performed on the publicly available Chaoyang colon histopathology dataset [24]. The dataset contains H&E-stained tissue patches from colonoscopy specimens obtained at Chaoyang Hospital, Beijing. These patches are categorized into four different classes: normal, serrated, adenocarcinoma and adenoma. Representative patches from each class are provided in Fig. 1.

All patches had an image size of 512 × 512 pixels. The training dataset consisted of 5,398 images; the validation dataset – 387 images; the testing dataset – 375 images. Stratification was done based on the class label for the splits in addition to slide-level grouping to ensure that there is no leakage of data from the training dataset to the other datasets. Images were normalized to the [–1, 1] range for training of the diffusion model; images were resized to 224 × 224 and ImageNet-normalized for embeddings extraction. During training random horizontal flips were used.

A generalized illustration of the proposed method is presented in Fig. 2. The method takes a histopathology reference patch as input and generates a synthetic image conditioned on the morphological features that are encoded by a pathology foundation model. Variation in the synthetic images arises from stochastic sampling. The method combines parameter-efficient adaptations of a pre-trained diffusion backbone with a learned embedding projector.

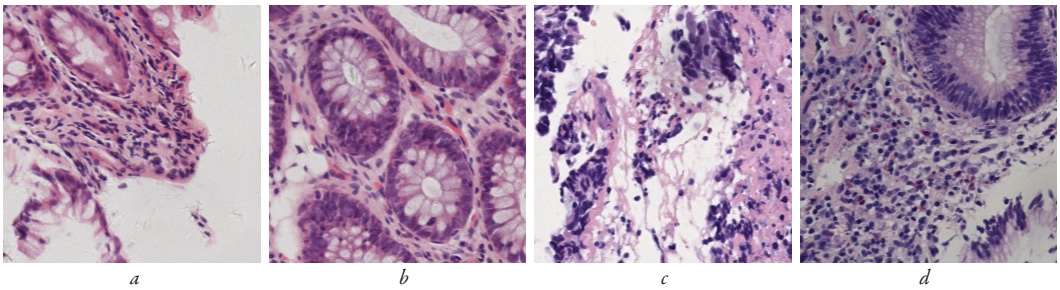
Proposed method builds upon stable diffusion 1.5 [9], which comprises a frozen variational autoencoder and a U-Net denoiser. The VAE provides 8× spatial compression, encoding 512 × 512 images into 64 × 64 latent representations with 4 channels. Let x denote an input image and $z_0 = E(x)$ its latent representation. The forward diffusion process adds Gaussian noise according to a predefined schedule. This process is described by formula

$$z_t = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{1}$$

where t indexes the diffusion timestep and $\bar{\alpha}_t$ denotes the cumulative noise schedule coefficient. The model is trained to predict the added noise. The loss function is described by formula

$$\mathcal{L} = \mathbb{E}_{x,\epsilon,t}\left[\left\|\epsilon - \epsilon_\theta(z_t, t, c)\right\|^2\right], \tag{2}$$

where c represents the conditioning signal and θ denotes the trainable parameters.



a *b* *c* *d*

Fig. 1. Representative histopathology patches from the Chaoyang colon dataset: *a* – normal; *b* – serrated; *c* – adenocarcinoma; *d* – adenoma

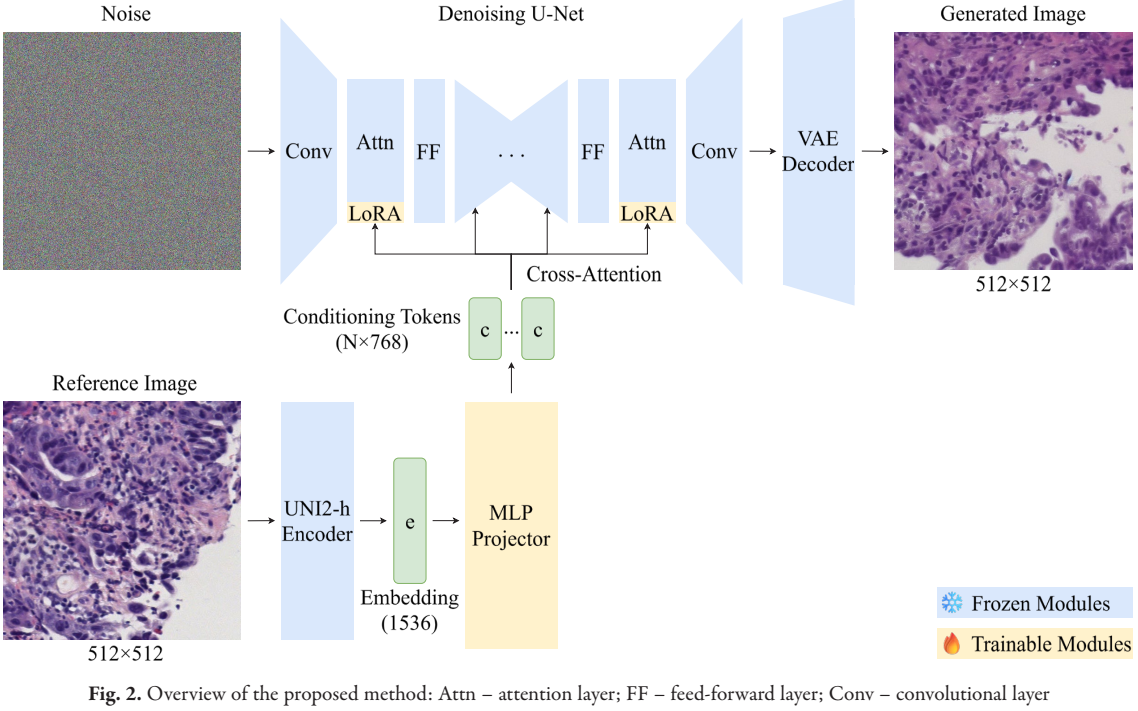


Fig. 2. Overview of the proposed method: Attn – attention layer; FF – feed-forward layer; Conv – convolutional layer

ISSN-L 2664-9969; E-ISSN 2706-5448

In the conditioning pathway, the original CLIP text encoder is replaced with the UNI2-h encoder. UNI2-h is a pathology foundational model based on ViT-H/14 architecture [11] that generates 1536-dimensional embeddings that have diagnostically relevant morphology. These embeddings are projected into the conditioning cross-attention space via a learnable projector. A projector was implemented as a multi-layer perceptron (MLP) with layer normalization and SiLU activation. The default configuration consists of two linear layers with a hidden dimension of 1024. The projector output is then reshaped into 4 conditioning tokens, each with a dimensionality of 768 (to match the cross-attention context dimension in the U-net). The transformation is described by formulas:

$$h=\sigma\left(LN\left(W_1 e+b_1\right)\right),\quad h\in\mathbb{R}^{d_h}, \tag{3}$$

$$c=\left[W_2 h+b_2\right]_{n_t\times d_c}\in\mathbb{R}^{n_t\times d_c}, \tag{4}$$

where e denotes the UNI2-h embedding, σ is the SiLU activation, and LN denotes layer normalization. The trainable null embedding has a shape of (1, 1536), which enables classifier-free guidance [25] during inference. For stable training the null embedding is initialized with very small random values (scale 0.01). The conditioning signal is dropped with a probability of 0.1 during training; this is done by replacing the null embedding before the projection with the null embedding. Thus, the dropout takes place at the embedding level and not after the embedding is projected. As such, the projector will learn to take into account both conditioned and unconditioned input via the same path. At test time, the guided predictions combine the conditional and unconditional predictions. This combination is described by formula

$$\tilde{\epsilon}=\epsilon_\theta(z_t,t,c_\varnothing)+w\cdot\left(\epsilon_\theta(z_t,t,c)-\epsilon_\theta(z_t,t,c_\varnothing)\right), \tag{5}$$

where w is the guidance scale and $c_\varnothing$ denotes the null-conditioned output.

Low-rank adaptation (LoRA) [15] is used on the query, key, value, and output projections of all U-Net attention layers, including both cross-attention and self-attention blocks. The adaptation modifies the weight matrices according to formula

$$W'=W+\frac{\alpha}{r}\cdot BA, \tag{6}$$

where W is a frozen pretrained weights matrix, $B\in R^{(d\times r)}$ and $A\in R^{(r\times d)}$ are trainable low-rank factor matrices with rank r and scalable factor α. In all of the experiments it is selected to set $\alpha = r$, or unit scale ($\alpha/r = 1$), thus ensuring that there is a consistent update magnitude across ranks. The original backbone's weights remain unchanged; the low-rank matrices and projectors are being trained.

To consolidate the above formulas into complete procedural descriptions, the training and inference procedures are presented as Algorithm 1 and Algorithm 2 (Fig. 3, 4).

The baseline configuration is summarized in Table 1.

1: procedure TRAIN$\left(D=\{(x_i,e_i)\},E,\epsilon_\theta,e_\varnothing,f_\varphi,p\right)$
2: repeat
3: $(x,e)\sim D$ ▷ Sample image – embedding pair
4: $z_0\leftarrow E(x)$ ▷ Encode to latent space
5: $\epsilon\sim\mathcal{N}(0,I),\ t\sim$ Uniform$(\{1,\dots,T\})$ ▷ Sample noise and timestep
6: $z_t\leftarrow\sqrt{\bar{\alpha}_t}\,z_0+\sqrt{1-\bar{\alpha}_t}\,\epsilon$ ▷ Forward diffusion
7: $m\sim$ Bernoulli(p) ▷ Per-sample CFG dropout
8: $\tilde{e}\leftarrow(1-m)\cdot e+m\cdot e_\varnothing$ ▷ Replace with null embedding
9: $c\leftarrow f_\varphi(\tilde{e})$ ▷ Project to cross-attention space
10: $\mathcal{L}\leftarrow\|\epsilon-\epsilon_\theta(z_t,t,c)\|^2$ ▷ Noise prediction loss
11: Update φ and $\{A, B\}$ ▷ Backbone W frozen
12: until converged

Fig. 3. Algorithm 1 training

1: procedure SAMPLE$\left(e,e_\varnothing,f_\varphi,\epsilon_\theta,w,T\right)$
2: $c\leftarrow f_\varphi(e),\quad c_\varnothing\leftarrow f_\varphi(e_\varnothing)$ ▷ Project both embeddings once
3: $z_T\sim\mathcal{N}(0,I)$ ▷ Sample initial noise
4: For $t = T, \dots, 1$ do ▷ Reverse diffusion over T steps
5: $\tilde{\epsilon}=\epsilon_\theta(z_t,t,c_\varnothing)+w\cdot(\epsilon_\theta(z_t,t,c)-\epsilon_\theta(z_t,t,c_\varnothing))$ ▷ Guided prediction
6: $z_{t-1}\leftarrow$ DenoisingStep$(z_t,\tilde{\epsilon},t)$ ▷ DDPM reverse update
7: return Decode(z_0) ▷ VAE decoder to pixel space

Fig. 4. Algorithm 2 inference

Table 1

Baseline model configuration

Module	Parameter	Value
LoRA	Rank/Alpha	8/8.0
	Target layers	Q, K, V, O (self- and cross-attention)
Projector	Hidden dimension	1024
	Depth	2 layers
	Activation	SiLU
	Normalization	LayerNorm
Conditioning	Number of tokens	4

Four architectural factors influence the total trainable parameter count: LoRA rank determines the size of low-rank matrices in each attention layer. The number of conditioning tokens, projector depth, and projector hidden dimension jointly determine the projector size. The trainable parameter breakdown for this baseline configuration is given in Table 2.

Table 2

Trainable parameters count for baseline model configuration

Component	Parameters count	Share
LoRA adapters	1,594,368	25.2%
Projector MLP	4,726,272	74.8%
Null embedding	1,536	<0.1%
Total trainable	6,322,176	100%
Stable diffusion 1.5 backbone (frozen)	859,520,964	–
Trainable/backbone	–	0.74%

To optimize the model AdamW was used with $\beta_1 = 0.9$ and $\beta_2 = 0.999$; in addition, differential weight decay was used (LoRA was set to $1e^{-3}$ in order to maintain pre-trained modifications minimal; the projector was set to $1e^{-4}$ to allow it to be flexible). Different learning rates of $5e^{-5}$ for LoRA adapter, and $5e^{-4}$ for the projector (a 10× baseline ratio) were used because the projector has to learn a new mapping

from scratch while LoRA is going to fine-tune the patterns already present. Ablation study investigates different ratios (1×, 5×, 20×) in order to determine the best compromise. All of the learning rates follow a cosine decay schedule and linear warm-up until about the first epoch. Training proceeded for 12 epochs with a batch size of 16 and BF16 mixed precision. Dropout of 0.05 was applied in the projector's MLP to regularize. In addition, to enable classifier-free guidance at inference, the conditioning signal was dropped with probability 0.1 during training by substituting a learned null embedding before projection.

Training requires around 12 GB of VRAM, so it fits on a 16 GB consumer GPU like NVIDIA's RTX 4080. Ablation experiments were run using cloud-based GPU access to shorten overall experimental run times. The implementation uses PyTorch 2.x along with HuggingFace's Diffusers for the diffusion backbone; custom modules for the projector and LoRA integration are used as well.

To measure how architectural design options affect the quality of generated images, systematic single-variable ablations were conducted in which one variable was changed from baseline (and all other variables were held constant). As the same optimization parameters were used for each ablation run to eliminate differences due to training procedures and isolate the effect of architectural design options, the resulting differences are directly related to the architecture.

The variants for the ablation factors and the ablation factors themselves can be found in Table 3. These were chosen due to the expected influence on the conditioning path and the adaptive capabilities. LoRA's rank controls an adapter's capacity with higher value increasing an adapter's expressiveness, however it may also result in overfitting. The number of tokens used for conditioning determines how granular the conditioning signal is sent through cross-attention. The projector's depth and hidden dimension will both affect the degree of non-linearity and capacity that exists within the embedding transformation. The learning rate ratio will affect the speed at which the projector adapts versus LoRA adapters.

Table 3

Ablation study factors and tested variants

Factor	Variants	Baseline
LoRA rank	4, 8, 16	8
LoRA alpha	matched to rank	8.0
Number of conditioning tokens	1, 2, 4, 8	4
Projector depth	2, 3 layers	2
Projector hidden dimension	512, 1024, 2048	1024
LR ratio (projector/LoRA)	1×, 5×, 10×, 20×	10×

The ablation study included 11 additional variant configurations for a total of 12 full training runs (each run lasted 12 epochs). DDPM scheduler was used at inference time with 30 denoising steps, since this configuration produced good image quality while keeping compute costs reasonable. To ensure reproducibility, deterministic seeds were used at generation time by computing the seed as follows: $seed[i] = base\_seed + i$. This produces the same set of samples every time that run is repeated regardless of batch size or number of GPUs used.

Early-stage studies suggested that guidance scales common in text-to-image models (i. e., 7.5) led to poor-quality images, which suggests that the UNI2-h embedding provides a stronger conditioning signal than the CLIP-derived text embeddings. Therefore, a search over guidance scales from 1.0 to 2.0 in steps of 0.1 was performed to find the best FID performance on the validation set. The optimal value is specified in the following section and fixed for all ablation comparisons.

The generation quality of the images was calculated through the use of FID and KID. Both FID and KID were computed using InceptionV3 feature space. A key drawback of the approach is that InceptionV3 has been pre-trained for natural images, not histopathology, however, it is well established that FID and KID can be applied for evaluating the quality of medical image synthesis and therefore allow for comparisons to be made with other works in the area. FID measures the distributional difference between true and generated images [26]. KID provides an unbiased estimator that performs particularly well when working with small sample sizes [27].

To find the best configuration, the validation FID was monitored on a per-epoch basis across all 12 training runs. Each validation image was encoded using UNI2-h as input to produce a single synthetic image under deterministic seeding. As such, this resulted in 144 (configuration, epoch) pairs, and the pair that had the lowest FID was used to determine both the optimal architecture for the model, and the checkpoint that would be carried forward into the final evaluation on the held-out test set.

Aggregate FID scores allowed to assess how closely the overall distributions of synthetic images matched original data. The results for per-class KID, along with their respective standard deviations helped provide insight into which classes were most difficult to reproduce. This may help guide further research and/or testing of the methods used in this research.

3. Results and Discussion

3.1. Parameter-efficient fine-tuning of the baseline model and inference configuration

The baseline model was trained for 12 epochs on the Chaoyang dataset. A validation FID was calculated through each epoch of the training process with a default guidance scale of 1.5 to identify the best checkpoint. In addition, as illustrated in Table 4, the FID score of the model declined rapidly early in training (from an initial value of 121.40 at epoch 1 to 82.46 by epoch 6). The lowest FID score (of 80.27) occurred at epoch 8. After this point, training continued to be oscillatory (i. e., FID rose to 101.40 by epoch 11 and then recovered to 84.74 at epoch 12). Therefore, the best checkpoint (epoch 8) was selected for use in subsequent experiments.

Following the choice of checkpoint, a sweep was performed on the epoch 8 checkpoint to select an appropriate guidance scale value for the best results. Preliminary experiments indicated that a guidance scale range of 7.0–10.0 used by standard text-to-image models resulted in artifacts when using the embedding-conditioned model, therefore, this sweep was limited to values from 1.0 to 2.0 in increments of 0.1.

Table 5 lists FID scores for evaluation sets using different guidance scales. The best guidance scale is 1.3 since it achieved a FID of 79.74. The FID did not exceed 82 from 1.3 to 1.8, therefore this range can be considered to be a flat optimum region. Performance decreased when either extreme was used (FID = 84.24 at a guidance scale of 1.0; FID = 84.29 at a guidance scale of 2.0). This combination of parameters (guidance scale 1.3; epoch 8 checkpoint; FID 79.74) will be used to compare to each of the ablations.

Table 4

Validation FID progression over training epochs (guidance scale 1.5)

Epoch	1	2	3	4	5	6	7	8	9	10	11	12
FID	121.40	105.47	98.33	92.11	83.64	82.46	85.71	80.27	92.26	86.04	101.40	84.74

ISSN-L 2664-9969; E-ISSN 2706-5448

Table 5

Validation FID as a function of classifier-free guidance scale (baseline epoch 8 checkpoint)

Guidance scale	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
FID	84.24	81.41	83.49	79.74	81.08	80.46	80.84	81.87	80.28	80.66	84.29

3.2. Systematic ablation study results

Each of the ablation variants underwent training for 12 epochs, as outlined previously. The LoRA alpha values are assigned the same value as their corresponding rank across each configuration. All of the configurations were evaluated at a guidance scale of 1.3 and the epoch that produced the lowest validation FID score was selected as the best performing epoch. The complete ablation results are presented in Table 6 sorted by FID in ascending order.

Table 6

Ablation study results (sorted by FID, ascending)

No.	LoRA rank	No. of tokens	Projector depth	Projector width	LR ratio	Best epoch	FID
1	4	4	2	1024	10×	12	77.59
2	8	4	2	1024	10×	8	79.74
3	8	4	2	1024	5×	6	79.94
4	8	4	2	1024	20×	5	80.64
5	8	8	2	1024	10×	12	80.87
6	16	4	2	1024	10×	6	82.98
7	8	4	2	2048	10×	5	83.57
8	8	4	3	1024	10×	8	83.66
9	8	2	2	1024	10×	5	83.69
10	8	4	2	512	10×	6	84.61
11	8	1	2	1024	10×	6	85.09
12	8	4	2	1024	1×	11	90.72

The lowest FID value was obtained in an ablation experiment using a LoRA rank of 4, which had a FID of 77.59 (an improvement of 2.7% compared to the baseline). The reduction of the rank from 8 to 4 resulted in a better performance, while an increase to 16 led to a decline in FID to 82.98. Notably, configuration with a rank of 16 and larger projector variants (width 2048, depth 3) achieved their maximum FID values at earlier epochs (5–8), while the rank 4 configuration continued improving through epoch 12.

The optimal amount for conditioning tokens for UNI2-h embedding representations is 4. The use of 1 and 2 tokens resulted in FID values of 85.09 and 83.69, which indicates the need for a greater number of tokens to fully represent the information contained within UNI2-h embeddings. When using 8 tokens, the FID value decreased to 80.87 (a slight drop from the baseline).

Projector architecture modifications did not perform better than the original. When changed from 2 layers to 3 layers, the FID was increased to 83.66. It was found that reducing the number of hidden dimensions to 512 decreased the FID to 84.61; however, an increase to 2048 produced a FID of 83.57. Therefore, the original 2 layer projector with 1024 hidden dimensions is still the best choice.

The learning rate ratio was found to have the most significant impact among the examined variables. The lowest performance, as indicated by an FID of 90.72, was obtained when the ratio was set to 1× (i. e., equal learning rates for the projector and LoRA). However, ratios of 5×, 10×, and 20× resulted in comparable performance (79.94, 79.74, and 80.64 respectively). This suggests that there is a threshold where the learning rate ratio has no further impact on performance above approximately 5×; below the threshold, the projector cannot effectively learn.

3.3. Optimal configuration, evaluation and discussion

The optimal configuration for ablation study used LoRA rank 4 and was otherwise identical to the base model for all other parameters. With this lower rank of 4 instead of 8 there are fewer trainable parameters for LoRA adapter layers (down from 1.59 million to 0.80 million). However, the same size projector of 4.73 million is retained. Therefore, there are now a total of 5.53 million trainable parameters (which represents 0.64% of the total) for the Stable Diffusion backbone. As a result, an increase of 2.7% in the validation FID score was achieved relative to the baseline model, while reducing total trainable parameters by 12.7%.

Evaluation of the best configuration was carried out on a held-out test set (375 images) at a guidance scale of 1.3. For comparison purposes both test set and validation set FIDs are included in Table 7. As expected, there is an increase in FID when comparing the validation split to the test split as distributions are likely to vary between the two.

Table 7

Validation and test set FID for optimal configuration

Split	Images	FID
Validation	387	77.59
Test	375	84.17

Table 8 presents per-class KID scores (×1000) on the test set alongside sample counts. The model achieved the best performance on adenoma (KID 13.51) despite it being the smallest class (33 samples), and worst on normal tissue (KID 48.67) which had 97 samples. This pattern suggests that generation quality is influenced more by within-class morphological variability than by sample size.

Table 8

Per-class KID scores (×1000) on test set

Class	Samples	KID (mean ± std)
adenoma	33	13.51 ± 8.14
adenocarcinoma	182	26.70 ± 13.18
serrated	63	28.18 ± 14.47
normal	97	48.67 ± 34.21

Representative synthetic images from Fig. 5 were created using the optimal configuration in addition to their respective conditioning input. The generated images preserved the architecture of the tissues as well as staining patterns for each of the four different classes. In conclusion, it was demonstrated that UNI2-h embeddings could be used as an alternative to text conditioning when generating histopathological data through efficient parameter adaptation.

The results from the ablation studies outlined in Table 6 demonstrated that the LoRA with rank of 4 provided the smallest validation FID of 77.59 when compared to the baseline rank-8 LoRA configuration, and performed about 2.7% better than the baseline. This result was somewhat surprising because, as may be expected based on the amount of modeling capacity afforded by higher-rank adaptations, higher-rank models were generally larger and therefore should have been capable of capturing more information. Although the rank-16 LoRA configuration and the extended projector variants (2048 hidden dimensions, 3 layers) reached their maximum FID score at an early (5–8) epoch. However, the LoRA with rank of 4 continued to improve up until the end of the

last epoch. A potential explanation is that lower-rank LoRA are advantageous due to a desirable bias-variance tradeoff for this dataset size; however, proving this will require a study of LoRA models trained on different-sized datasets.

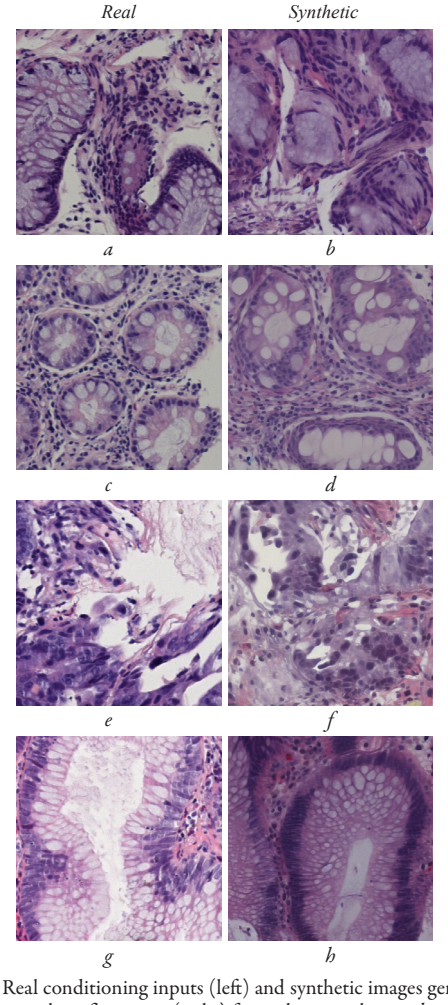


Fig. 5. Real conditioning inputs (left) and synthetic images generated by the optimal configuration (right) for each tissue class: *a, b* – normal; *c, d* – serrated; *e, f* – adenocarcinoma; g, *h* – adenoma

Visual comparison of conditioning input and output from UNI2-h suggests that it is capturing a level of morphological detail in the embedding space that is useful enough to drive image generation (including architectural and staining details) for all 4 tissue types shown. However, visual inspection alone cannot establish diagnostic fidelity, which would require evaluation by domain experts or downstream task validation.

Compared to the text-conditioned approach explored in the prior work [19], which achieved per-class FID scores ranging from 113 to 138 on the same dataset, the embedding-conditioned method achieved substantially lower FID values. This improvement may be attributed to the replacement of CLIP-based text conditioning with pathology foundation model embeddings that were trained on large whole-slide image collections using self-supervised objectives specifically designed for histopathology.

Ablation results show that when the learning rates for both projector and LoRA are the same, there is a significant reduction in performance (FID 90.72) compared with the other ratios of 5× or greater (FID 79.74–80.64 for 5×, 10×, and 20×, respectively). Therefore, it appears that a learning rate for the projector that is much larger than those used by LoRA adapter will be needed for the projector to successfully map from the UNI2-h embedding space to the cross-attention conditioning space. However, the precise minimum ratio at which this transition occurs was not determined and would require finer-grained ablation in future work.

The benefits of the proposed approach are based upon combining parameter-efficient adaptation with direct foundation model conditioning as an alternative to both fully fine-tuned approaches and text-dependent approaches. In terms of computation, LRDM [13] trained the diffusion backbone from scratch and used 6 RTX 8000 GPUs with 15 million training patches, while proposed method uses only 5.53 million trainable parameters (0.64% of the stable diffusion backbone) and has approximately 12 GB of GPU memory available. Although the fair comparison of the two is not possible due to the differences in dataset, tissue type, and assessment methodology, the substantial decrease in processing resources provides a tangible advantage for deploying in resource-constrained environments. As to CytoDiff [14], it does use parameter-efficient LoRA adaptation and therefore does not have to fine-tune the entire model, it also continues to rely on the CLIP text encoder, while proposed method removes the text-conditioned path completely and receives the condition signal from the morphological characteristics of the tissue, thus eliminating the need for the user to design prompts.

Special features of research results are:

– it is shown that embeddings from a pathology foundation model can be used as an alternative to text conditioning in generating synthetic images in histopathology without having to use either textual supervisory information or prompt engineering;

– it is found through extensive systematic ablation studies that the lower LoRA ranks (i. e., rank = 4 instead of 8 or 16), result in improved performance when tested on moderately sized histopathology data sets;

– it is established that the learning rate ratio between projector and LoRA components substantially affects training dynamics, with equal learning rates leading to degraded convergence.

The practical significance of this research findings lies in an opportunity to use synthetic image generation for enhancing (augmenting) data used in a classification task using histopathology images. Previously, it was shown through the use of GANs in digital pathology that augmenting with generated images would enhance classifier accuracy when training classifiers for less represented types of tissues (classes) [6]. Additionally, as a result of the parameter efficiency of the method proposed here, such augmentation strategies may be possible in clinical environments with limited availability of high-performance computing. Removing the need for a text prompt also eliminated the need for manual authorship of captions for the text; this is both impractical and unscalable for histopathology patch images that rarely have a meaningful textual description.

On the other hand, there is a number of limitations of research:

– all experiments were carried out on one dataset (Chaoyang colon histopathology) so the research results can't be generalized to other types of tissues and staining protocols;

– the fixed resolution of 512 × 512 pixels corresponds to the original training resolution of Stable Diffusion 1.5 but it could be suboptimal for many applications of histopathology requiring high-resolution patches or multi-scale generation;

– evaluation was done by means of the FID and KID metrics calculated in the feature space of InceptionV3, which was trained on natural images instead of histopathology images. Although these metrics have become widely used as reference in the field of medical image synthesis, they don't fully reflect differences in diagnostic quality;

– downstream classification performance wasn't evaluated in this research, which limits the assessment of practical utility for data augmentation applications.

The disadvantages of research are the lack of validation on multiple datasets with different types of tissues and no evaluation of the utility of

ISSN-L 2664-9969; E-ISSN 2706-5448

synthetic images for the downstream tasks to confirm their usefulness for classifier training.

Further research may include the expansion of the evaluation to more histopathology datasets of different organs and staining methods, the investigation of multi-scale generation techniques to generate whole-slide images, and evaluating the use of synthetic images as a means to improve the accuracy of classifiers by using them to augment the training set.

4. Conclusions

1. The embedding-to-cross-attention projector architecture was designed as a two-layer MLP with LayerNorm and SiLU activation that converts UNI2-h pathology foundation model embeddings into 4 conditioning tokens to use within the U-Net cross-attention space. In combination with low-rank adaptation (LoRA) applied to each of the self-and cross-attention layers, this architecture was able to be trained with just 6.32 million trainable parameters in the baseline configuration (0.74% of the 859.5 million Stable Diffusion 1.5 backbone), while the remaining weights were frozen. The potential of using pathology foundation models to condition and embedding was validated through the model being able to generate images for all four tissue types (normal, serrated, adenocarcinoma, adenoma) without any text prompt. The results demonstrated that the embedding conditioned method produced significantly lower FID values compared to the text conditioned approach developed in prior work [19], which produced FID scores ranging from 113–138 on the same dataset, and thus UNI2-h embeddings are a more powerful and morphologically richer conditioning signal than CLIP-based text encodings. As such, this result enables the creation of synthetic histopathological images without the need for manually creating captions for individual patches of images which can be impractical due to the lack of meaningful textual descriptions.

2. A systematic ablation analysis was performed across 12 different configurations where five architectural variables were each changed individually and all other parameters remained the same. The parameters that were tested included LoRA rank (4, 8, 16), the number of conditioning tokens (1, 2, 4, 8), projector depth (2, 3), projector hidden dimensionality (512, 1024, 2048) and learning rate ratio between the projector and LoRA adapters (1×, 5×, 10×, 20×). Learning rate ratio demonstrated the greatest influence among all factors as equal learning rates (1×) produced an FID score of 90.72, while all higher ratios (5×, 10×, 20×) produced similar performance (FID scores ranged from 79.74 to 80.64) indicating a threshold at which point the projector is unable to effectively learn the embedding-to-cross-attention mapping. The optimal number of conditioning tokens was identified as 4 (FID 79.74) while fewer numbers of conditioning tokens (i. e., 1 token: FID 85.09, 2 tokens: FID 83.69) and greater numbers of conditioning tokens (i. e., 8 tokens: FID 80.87) produced lower quality results. None of the additional modifications made to the projector (i. e., 3 layers: FID 83.66, 512 hidden dimension: FID 84.61, 2048 hidden dimension: FID 83.57) improved upon the baseline two-layer 1024-dimension configuration. Thus, these results provide empirical evidence for practitioners by showing that the learning rate required for effective training of the conditioning pathway must be calibrated carefully relative to the LoRA adapters, and that neither increasing the depth nor the width of the projector will result in significantly better performance on this dataset.

3. Optimal configuration was determined to be LoRA rank 4, with every other parameter equal to those of the baseline (i. e., 4 conditioning tokens; 2-layer projector; 1024-dimension hidden layers; 10× learning rate ratio). It achieved an optimal validation FID score of 77.59 – a 2.7% improvement over the baseline LoRA rank-8 configuration. At the same time the number of trainable parameters was reduced by 12.7%, resulting in a total of 5.53 million (0.64% of the backbone) trainable parameters. A held-out test set consisting of 375 images resulted in an FID score of 84.17. Within class KID scores were between 13.51 for adenoma and 48.67 for normal tissue indicating that the quality of the generated images is more affected by the within-class morphologic variation than by the number of samples. Approximately 12 GB of GPU memory are required for this method to operate. This will enable the deployment of this method to consumer-grade GPUs (for example, NVIDIA RTX 4080 with 16 GB VRAM) for histopathology data augmentation when textual annotation details are not available in resource-constrained environments.

Conflict of interest

The authors declare that they have no conflicts of interest in relation to the current research, including financial, personal, authorship, or any other, that could affect the research, as well as the results reported in this paper.

Financing

The research was conducted without financial support.

Data availability

The Chaoyang colon histopathology dataset used in this research is publicly available [24].

Use of artificial intelligence

The authors used artificial intelligence technologies within acceptable limits to provide their own verified data, which is described in the research methodology section.

Authors' contributions

Sergii Kuzmin: Conceptualization, Methodology, Software, Investigation, Writing – original draft; ***Oleh Berezsky***: Conceptualization, Validation, Writing – review and editing, Supervision.

References

1. Litjens, G., Bandi, P., Ehteshami Bejnordi, B., Geessink, O., Balkenhol, M., Bult, P. et al. (2018). 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. *GigaScience, 7 (6).* https://doi.org/10.1093/gigascience/giy065
2. Walsh, E., Orsi, N. M. (2024). The current troubled state of the global pathology workforce: a concise review. *Diagnostic Pathology, 19 (1).* https://doi.org/10.1186/s13000-024-01590-2
3. Guan, H., Yap, P.-T., Bozoki, A., Liu, M. (2024). Federated learning for medical image analysis: A survey. *Pattern Recognition, 151,* 110424. https://doi.org/10.1016/j.patcog.2024.110424
4. Zhang, Y., Kang, B., Hooi, B., Yan, S., Feng, J. (2023). Deep Long-Tailed Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45 (9),* 10795–10816. https://doi.org/10.1109/tpami.2023.3268118
5. Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K. J. et al. (2019). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine, 25 (8),* 1301–1309. https://doi.org/10.1038/s41591-019-0508-1
6. Jose, L., Liu, S., Russo, C., Nadort, A., Di Ieva, A. (2021). Generative Adversarial Networks in Digital Pathology and Histopathological Image Processing: A Review. *Journal of Pathology Informatics, 12 (1),* 43. https://doi.org/10.4103/jpi.jpi_103_20
7. Saad, M. M., O'Reilly, R., Rehmani, M. H. (2024). A survey on training challenges in generative adversarial networks for biomedical image analysis. *Artificial Intelligence Review, 57 (2).* https://doi.org/10.1007/s10462-023-10624-y
8. Dhariwal, P., Nichol, A. (2021). Diffusion models beat GANs on image synthesis. *arXiv:2105.05233.* https://doi.org/10.48550/arXiv.2105.05233
9. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B. (2022). High-Resolution Image Synthesis with Latent Diffusion Models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* 10674–10685. https://doi.org/10.1109/cvpr52688.2022.01042

10. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S. et al. (2021). Learning transferable visual models from natural language supervision. *arXiv:2103.00020.* https://doi.org/10.48550/arXiv.2103.00020

11. Chen, R. J., Ding, T., Lu, M. Y., Williamson, D. F. K., Jaume, G., Song, A. H. et al. (2024). Towards a general-purpose foundation model for computational pathology. *Nature Medicine, 30,* 850–862. https://doi.org/10.1038/s41591-024-02857-3

12. Yellapragada, S., Graikos, A., Prasanna, P., Kurc, T., Saltz, J., Samaras, D. (2024). PathLDM: Text conditioned Latent Diffusion Model for Histopathology. *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV),* 5170–5179. https://doi.org/10.1109/wacv57701.2024.00510

13. Graikos, A., Yellapragada, S., Le, M.-Q., Kapse, S., Prasanna, P., Saltz, J., Samaras, D. (2024). Learned Representation-Guided Diffusion Models for Large-Image Generation. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* 8532–8542. https://doi.org/10.1109/cvpr52733.2024.00815

14. Boada, J. C., Umer, R. M., Marr, C. (2025). CytoDiff: AI-Driven Cytomorphology Image Synthesis for Medical Diagnostics. *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW),* 1136–1144. https://doi.org/10.1109/iccvw69036.2025.00122

15. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S. et al. (2022). LoRA: low-rank adaptation of large language models. *arXiv:2106.09685.* https://doi.org/10.48550/arXiv.2106.09685

16. Ho, J., Jain, A., Abbeel, P. (2020). Denoising diffusion probabilistic models. *Proceedings of the 34th International Conference on Neural Information Processing Systems, 34,* 6840–6851. https://doi.org/10.48550/arXiv.2006.11239

17. Yellapragada, S., Graikos, A., Triaridis, K., Prasanna, P., Gupta, R., Saltz, J., Samaras, D. (2025). ZoomLDM: Latent Diffusion Model for multi-scale image generation. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* 23453–23463. https://doi.org/10.1109/cvpr52734.2025.02184

18. Mao, Y., Li, H., Pang, W., Papanastasiou, G., Yang, G., Wang, C. (2024). SeLoRA: self-expanding low-rank adaptation of latent diffusion model for medical image synthesis. *arXiv:2408.07196.* https://doi.org/10.48550/arXiv.2408.07196

19. Berezsky, O., Melnyk, G., Liashchynskyi, P., Pitsun, O.; Babichev, S., Lytvynenko, V. (Eds.) (2025). Biomedical Image Datasets. *Lecture Notes on Data Engineering and Communications Technologies, vol 244.* Cham: Springer, 61–82. https://doi.org/10.1007/978-3-031-88483-2_3

20. Berezsky, O., Liashchynskyi, P., Melnyk, G., Dombrovskyi, M., Berezkyi, M. (2024). Synthesis of biomedical images based on generative intelligence tools. *Proceedings of the 7th International Conference on Informatics & Data-Driven Medicine (IDDM 2024). Birmingham. CEUR Workshop Proceedings, 3892,* 349–362. Available at: https://ceur-ws.org/Vol-3892/paper23.pdf

21. Berezsky, O., Liashchynskyi, P., Pitsun, O., Izonin, I. (2024). Synthesis of Convolutional Neural Network architectures for biomedical image classification. *Biomedical Signal Processing and Control, 95,* 106325. https://doi.org/10.1016/j.bspc.2024.106325

22. Berezsky, O., Liashchynskyi, P., Pitsun, O., Melnyk, G. (2024). Method and Software Tool for Generating Artificial Databases of Biomedical Images Based on Deep Neural Networks. *6th International Conference on Informatics & Data-Driven Medicine Bratislava.* https://doi.org/10.48550/arXiv.2405.16119

23. Kuzmin, S., Berezsky, O. (2025). Analysis of diffusion models and biomedical image generation tools. *Computer Systems and Information Technologies, 2,* 8–19. https://doi.org/10.31891/csit-2025-2-1

24. Zhu, C., Chen, W., Peng, T., Wang, Y., Jin, M. (2022). Hard Sample Aware Noise Robust Learning for Histopathology Image Classification. *IEEE Transactions on Medical Imaging, 41 (4),* 881–894. https://doi.org/10.1109/tmi.2021.3125459

25. Ho, J., Salimans, T. (2022). *Classifier-free diffusion guidance.* arXiv:2207.12598. https://doi.org/10.48550/arXiv.2207.12598

26. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *31st Conference on Neural Information Processing Systems (NIPS 2017).* Long Beac. https://doi.org/10.48550/arXiv.1706.08500

27. Bińkowski, M., Sutherland, D. J., Arbel, M., Gretton, A. (2018). Demystifying MMD GANs. *International Conference on Learning Representations (ICLR).* https://doi.org/10.48550/arXiv.1801.01401

✉ ***Sergii Kuzmin,*** *PhD Student, Department of Automated Control Systems, Lviv Polytechnic National University, Lviv, Ukraine, e-mail: kuzminos22@gmail.com, ORCID: https://orcid.org/0009-0001-7182-2883*

Oleh Berezsky, *Doctor of Technical Sciences, Professor, Department of Computer Engineering, West Ukrainian National University, Ternopil, Ukraine, ORCID: https://orcid.org/0000-0001-9931-4154*

✉ *Corresponding author*