Scientific Reports

https://doi.org/10.1038/s41598-026-49457-4

Article in Press

Diverse image generation with diffusion models and cross class label learning for polyp classification

Received: 22 September 2025

Accepted: 15 April 2026

Published online: 18 May 2026

Cite this article as: Sharma V., Jha D., Bhuyan M.K. *et al.* Diverse image generation with diffusion models and cross class label learning for polyp classification. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-49457-4

Vanshali Sharma, Debesh Jha, M. K. Bhuyan, Pradip K. Das & Ulas Bagci

Diverse Image Generation with Diffusion Models and Cross Class Label Learning for Polyp Classification

Vanshali Sharma[1,*], Debesh Jha[2], M.K. Bhuyan[3], Pradip K. Das[1], and Ulas Bagci[2]

[1]Department of Computer Science & Engineering, Indian Institute of Technology Guwahati, Guwahati, 781039, India
[3]Department of Electronics & Electrical Engineering, Indian Institute of Technology Guwahati, Guwahati, 781039 , India
[2]Machine & Hybrid Intelligence Lab, Department of Radiology, Northwestern University, Chicago, 60611, USA
[*]Correspondence should be addressed to Vanshali Sharma at vanshalisharma@alumni.iitg.ac.in.

ABSTRACT

Pathologic diagnosis is a critical phase in deciding the optimal treatment procedure for dealing with colorectal cancer (CRC). Colonic polyps, precursors to CRC, can pathologically be classified into two major types: adenomatous (malignant potential) and hyperplastic (benign). Various imaging techniques, such as narrow band imaging (NBI) and white light imaging (WLI), are adopted in capturing polyp-specific features for accurate classification and have different advantages. However, the existing classification techniques mainly rely on a single imaging modality and show limited performance due to data scarcity. Recently, generative artificial intelligence has been gaining prominence in overcoming such issues, especially with various generation-controlling mechanisms using text prompts and images. However, such mechanisms require class labels to make the model respond efficiently to the provided control input. In the colonoscopy domain, such controlling mechanisms are rarely explored; specifically, the text prompt is a completely uninvestigated area. Moreover, the unavailability of expensive class-wise labels for diverse sets of images limits such explorations. This raises the key question of how diverse and clinically meaningful colonoscopy images can be generated in a text-controlled manner from limited annotated data. Therefore, in this work, we develop a novel model, ***PathoPolyp-Diff***, that generates text-controlled synthetic images with diverse characteristics in terms of pathology, imaging modalities, and quality, enabling more effective augmentation of downstream diagnostic models. The proposed model follows a two-stage process: first, the model learns to distinguish polyp from non-polyp characteristics, and then it focuses on pathology-specific features. In the process, we introduce cross-class label learning to make the model learn features from other classes, reducing the burdensome task of data annotation. We validate the effectiveness of text-controlled synthesis and cross-class label learning by performing polyp classification (adenomatous/hyperplastic) with different imaging modalities (NBI/WLI) and text prompts. The experimental results show that incorporating the proposed synthetic images for data augmentation yields an improvement of up to 7.91% in balanced accuracy on a publicly available dataset, highlighting the utility of our approach for enhancing downstream classification performance. Moreover, cross-class label learning achieves a statistically significant improvement of up to 18.33% in balanced accuracy during video-level analysis. The code is available at https://github.com/Vanshali/PathoPolyp-Diff.

Introduction

Colorectal cancer (CRC) is the second most common cause of cancer-related mortality, accounting for about 1 million deaths per year[1]. The precursor lesions of CRC occur in the form of polyps, which in their initial stages are likely to be non-cancerous. However, they could turn into invasive cancer over time. The risk associated with such abnormal growth depends on the pathological characteristics of the lesion. For instance, a hyperplastic polyp is categorized as benign and generally not a cause of concern, whereas an adenomatous polyp might have the potential to become cancerous. Hence, early detection and categorization of polyps are necessary for optimal surgical procedures and effective treatment. Colonoscopy is a gold standard tool widely adopted for this purpose, in which a camera-mounted colonoscope is inserted into the patient's colon to investigate it for any abnormalities or lesions. This procedure encompasses various optical imaging technologies to capture polyp-specific features. Some such imaging techniques include white light imaging (WLI) and narrow band imaging (NBI). Among these imaging modalities, WLI is the most commonly used technique; however, it fails to capture topographic contrast, such as polyp elevation and pit patterns that increase the lesion miss rate. On the contrary, NBI uses electronically activated filters to limit the wavelengths of red, green and blue lights, which helps accentuate the superficial mucosa and vascular details[2]. Despite its advantageous applicability in differentiating benign and neoplastic polyps, a standard colonoscopy procedure follows WLI. Besides clinical use, the research community is mainly dependent on images captured

using a single imaging technique, i.e. WLI. The reason is the unavailability of publicly accessible image data encompassing different imaging modalities and

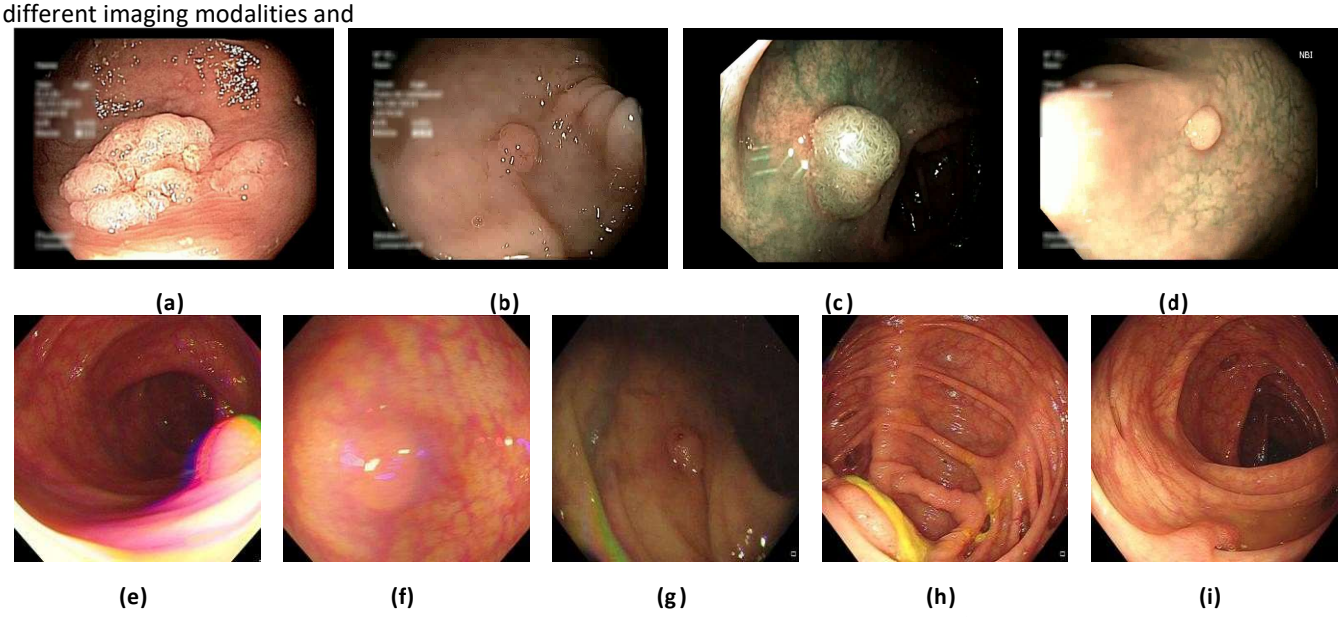


Figure 1. Sample images from the ISIT-UMR[9] and SUN Database[10,11] depicting (a) adenomatous polyp in WLI (b) hyperplastic polyp in WLI, (c) adenomatous polyp in NBI, (d) hyperplastic polyp in NBI, (e)-(h) represent low-quality frames with artefacts, namely, ghost colors, motion blur, low illumination, and fecal depositions, respectively, and (i) shows a good-quality polyp image.

pathological conditions.

The recent breakthrough in generative artificial intelligence (AI) has sparked the potential to obtain realistic medical images spanning diverse sets of classes. Consequently, various tasks in the medical domain, including super-resolution[3], image reconstruction[4], segmentation[5], and image-to-image translation[6], have witnessed remarkable performance enhancements using generative adversarial network (GAN) based methodologies. GANs are one the prominent types of generative AI techniques; however, they suffer from convergence instability and show limited contributions and control over data generation. The advent of diffusion models marked a significant stride in dealing with such issues. These models facilitate a better control and manipulation ability to generate diverse, realistic images. This control can be achieved using conditioning in the form of text, images, or both. More recently, numerous works[7,8] in different fields focused on text-to-image synthesis have been proposed. Such approaches use random noise and text prompts as input to produce synthetic images. These images are expected to represent the conditions provided in the text. The literature utilizes existing text-to-image techniques by fine-tuning foundation models on the downstream tasks. This fine-tuning process requires text prompts that describe the desired synthetic images appropriately.

In the colonoscopy domain, diffusion models have not been explored much, especially with different conditioning possibilities. Some of the existing works[12,13] mainly focus on generating polyp images with binary mask conditioning. These limited explorations render the study of pathological conditions (adenoma/hyperplastic) of polyps and various imaging modalities unexplored. Additionally, the significance of controlling polyp generation using text prompts remains overlooked. Moreover, a well-designed dataset with comprehensive annotations is essential for training generative models. This ensures their ability to grasp intricate patterns linked to polyp-characterizing features and their associated pathological conditions. However, obtaining labels for each subtask, considering pathology, quality, and imaging modality, could be significantly expensive. Therefore, in this framework, we develop a method to perform cross-class label learning that helps leverage annotations from other classes and produce synthetic images representing a combination of text prompts. This approach ensures obtaining synthetic polyp images with different pathological characteristics (adenoma/hyperplastic) captured with different imaging techniques (WLI/NBI). This diverse set is obtained while maintaining the quality, overcoming the general artifacts present in a colonoscopy video: ghost colors, motion blur, low illumination, and fecal deposition. Some sample images categorized under the aforementioned classes are shown in Fig. 1. Our contributions can be summarized as:

- Generated diverse set of colonoscopy images using text-controlled mechanism: We develop PathoPolyp-Diff, a novel model to generate text-controlled synthetic images that cover a wide range of categories, including pathology, imaging modalities, and quality. It can generate adenomatous and hyperplastic polyps combined with the desired imaging modalities, including NBI and WLI. To the best of our knowledge, this is the first work to explore text prompts to generate synthetic colonoscopy images.

- Introduced the concept of cross-class label learning: We introduce the concept of cross-class label learning that allows the model to learn labels from different classes, hence expanding the diversity of data generation and reducing the cumbersome task of dataset annotation. Additionally, we integrate this learning with a weighted mechanism and study the impact of weights with text prompts.

- Improved polyp classification performance: We report enhanced polyp classification outcomes when PathoPolyp-Diff's generated images are used to augment the real pathological dataset. This enhancement is achieved without additional manual annotations. We also performed an exhaustive analysis to study the impact of different text prompts on the quality of synthetic images generated for improved polyp classification.

Related Works

With the advent of generative AI, several works have been proposed to generate synthetic colonoscopy images. The main focus of existing related works has been to perform synthetic polyp generation irrespective of the pathological characteristics of polyps and the imaging modality. The techniques adopted so far can broadly be divided into two categories, namely, GAN, and Diffusion Models.

Generative Adversarial Network (GAN) based Techniques

The initial frameworks for polyp generation are based on the adversarial concept and adopt different variants of GANs. For example, Shin et al.[14] used a conditional-GAN approach to translate normal colonoscopy images to polyp images. This translation is achieved using an input-conditioned image which is a combination of an edge map and a polyp binary mask. A similar concept of converting normal frames to polyp frames is proposed in[15]. They utilized a conditional GAN architecture to produce polyps with varied characteristics by controlling the input-conditioned binary mask values. Such conditional translation is also reported by Fagereng et al.[16]. They developed a framework called PolypConnect which uses an EdgeConnect model to convert clean colon images to polyps when given an edge map and a polyp mask. Sasmal et al.[17] performed polyp generation using DCGAN and used the obtained synthetic polyps to enhance classifier performance for differentiating adenoma and hyperplastic polyps. An identical augmentation approach is followed by Adjei et al.[18] using synthetic polyps generated using a Pix2Pix model. Unlike the traditional GAN architecture, He et al.[19] introduced an attacker in the framework to obtain false negative images. Sams and Shomee[20] utilized a StyleGAN2-ada to generate random binary masks, which are combined with colon images. This integrated image is used as an input for a conditional GAN to obtain synthetic polyp images. The above methods focused on polyp generation irrespective of the imaging modalities. However, a few works have used GAN-based approaches to transfer styles between different imaging modalities, such as WLI and NBI. Golhar et al.[21] utilized the GAN inversion approach, which uses a latent representation of images to perform translation between NBI and WLI modalities. Following this technique, interpolation methods are used to change the polyp size. Similarly, Bhamre et al.[22] used CycleGAN to convert WLI images to NBI images. Although these few existing works established the significance of NBI images over WLI images in polyp classification, the generation of new synthetic polyp images with different imaging modalities has not been explored in the literature.

Diffusion Model based Techniques

The related literature involves only a few works focused on polyp image generation. Machacek et al.[12] used a conditional diffusion probabilistic model to produce synthetic polyp images using synthetic masks. They validated the effectiveness of generated data by utilizing it for training polyp segmentation models. Pishva et al.[13] performed polyp generation using two diffusion models. The two models are fine-tuned on cropped-out polyps and clean colon images, respectively. This fine-tuning is followed by performing an inpainting using the latter model and cropped-out images. Du et al.[23] proposed an adaptive refinement semantic diffusion model which considers the polyp and background ratio to adjust the diffusion loss. They also incorporated a pre-trained segmentation model that modifies the refinement loss depending on the difference between the predicted mask of the synthetic polyp and the binary mask used for its generation. The above-mentioned approaches

followed a similar pattern of polyp generation using binary masks. The impact of text prompt based training and the generation of colonoscopy images with different imaging modalities still remains unexplored.

Methodology

An overview of our approach is illustrated in Fig. 2. Our novel framework, referred to as PathoPolyp-Diff, utilizes dual-stage training. The two stages, *Stage-I* (Feature Distillation) and *Stage-II* (Pathological Synthesis), aim at generating colonoscopy images with diverse polyp types in different imaging modalities. They perform complementary tasks and the difference between the two lies in their training process. The *Stage-I* network distils knowledge into the *Stage-II* model in the form of a large set of features that enables it to differentiate between polyp and non-polyp characteristics and further helps to generate images for cross-class labels. Complementary to it, the training process in *Stage-II* allows the model to learn pathological details and different imaging modality-related patterns. More specifically, by structuring the method in two stages, we allow the model to first capture fundamental features, for which more samples are available, before learning the specific pathological classes,

which have relatively fewer samples.

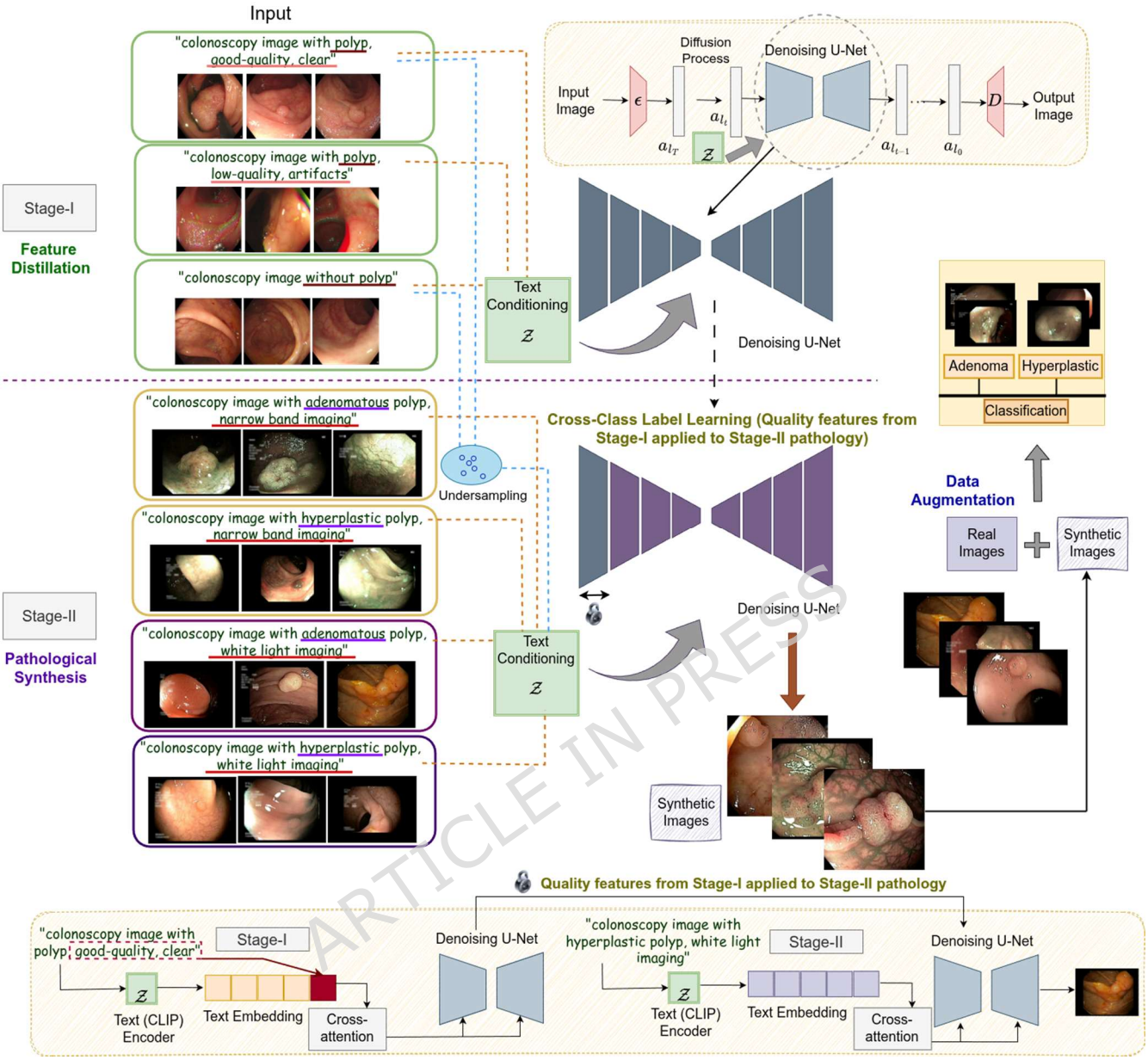


Figure 2. Overview of the proposed framework. It consists of two stages and uses various text conditioning to control the generation process. In Stage-II, some undersampled data from Stage-I is used for a smoother learning process. Also, the first block of U-Net is kept locked in the second stage. The performance of the proposed model is validated using a classification process which uses a combination of real and synthetic images in different proportions. The real images are taken from the ISIT-UMR[9] and SUN Database[10,11].

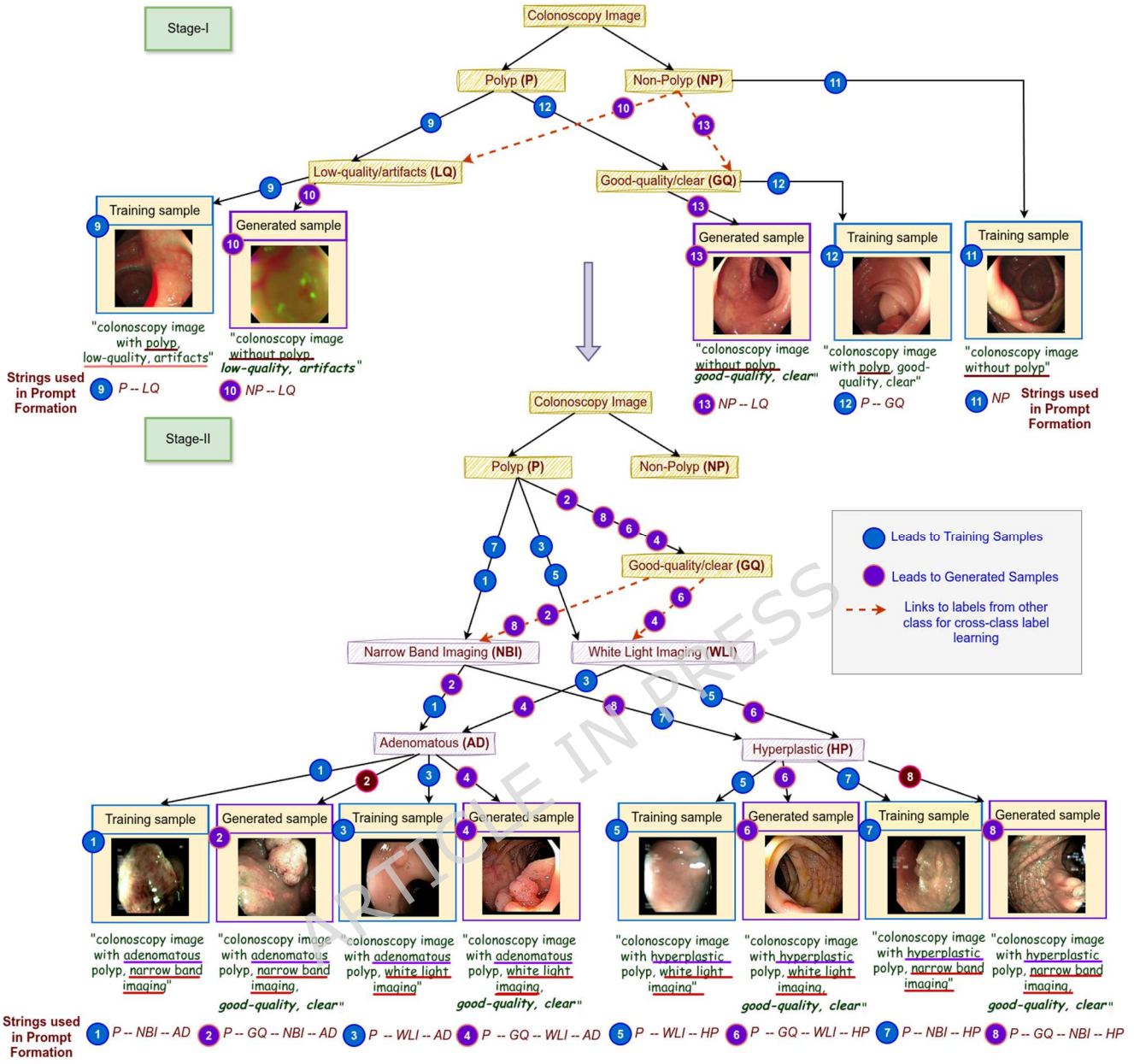


Figure 3. Flowchart depicting the different combinations of text prompts and cross-class labels used to generate images. The yellow nodes represent levels 1-3, while the nodes in another color represent levels 4-5. The solid arrows denote the labels already present in the dataset, whereas the dashed arrows represent the labels learnt from other classes (cross-class labels). Each number on a solid/dashed line represents the combination of strings used to form tokens for text prompts used in training/inference. For instance, following number *'8'*, we obtain the text prompt *"colonoscopy image with a hyperplastic polyp, narrow band imaging, good quality, clear"*, where *"good quality, clear"* are part of indirectly inferred tokens and other are already present in the training annotations. The training sample images are taken from ISIT-UMR[9] and SUN Database[10,11].

Architectural Details

Our approach utilizes the stable diffusion (SD) model's[24] architecture with different settings and transfer learning approach. SD is a text-based image generation network that works on the principle of latent diffusion model (LDM). It is introduced to circumvent the issue of the high computational requirements of standard diffusion models (DM)[25]. This improvement is

achieved by executing the diffusion process on latent space instead of pixel level using an autoencoding procedure. To understand the advantages of using LDM over DM, some basic details about the two concepts are given below.

Standard DMs: The standard DMs follow a parameterized backward process of a fixed Markov Chain to gradually denoise a noisy image a_t. It acts as a sequence of autoencoders $\varepsilon_\vartheta(a_t,t)$ that serves as a denoising framework to predict the denoised version of a_t. Here, t is uniformly sampled between $[1,T]$, and T denotes the noise steps. The related objective can be defined in a refined form as:

$$L_{DM} := \mathbb{E}_{a,\varepsilon,t}[\|\varepsilon - \varepsilon_\theta(a_t,t)\|_2^2] \quad (1)$$

where $\varepsilon \sim N(0,1)$.

Standard LDMs: These models leverage the ability of encoder and decoder architectures to represent significant information in compressed form and reconstruct it back in its original form. Such an attempt to use latent space enables the model to focus on important semantic details and perform efficiently with low computational resources. LDMs also keep track of time steps t_i and are embedded with the U-Net architecture. However, instead of using an image a directly, it is processed through an encoder E to obtain a_l.

$$L_{LDM} := \mathbb{E}_{E(a),\varepsilon,t}[\|\varepsilon - \varepsilon_\theta(a_{l_t},t)\|_2^2] \quad (2)$$

SD has three main components, namely, a variational autoencoder (VAE)[26], a U-Net[27], and a text encoder, called CLIP[28]. The VAE comprises an encoder E that converts the input image $a \in R^{H\times W\times 3}$ in a latent space for low resource overhead during complete processing and a decoder D that translates the encoded representation into a reconstructed image a. SD follows two processes similar to a standard LDM, i.e., forward diffusion and reverse diffusion. During forward diffusion, Gaussian noise is iteratively introduced in the image and during the reverse diffusion step, a U-Net architecture with ResNet blocks is employed to obtain denoised outcomes. This step is supported by text embeddings which represent the textual information about the image in a numeric form. These text embeddings are obtained using a pre-trained CLIP model that acts as conditioning on the image reconstruction.

Conditioning Mechanism: DMs, like any other generative model, are capable of conditioning the sampling to guide the generation process. It is achieved by modeling $p(a|b)$ where b denotes the condition that could be in the form of text or image embeddings. The SD model incorporates this concept in a more effective way by using a cross-attention mechanism. This mechanism is used to fuse text embeddings with the U-Net structure after pre-processing b through a text encoder Z (in case of text prompts). It encodes b into an intermediate representation, which is then mapped to U-Net architecture using cross-attention layers. This conditioning is integrated with the overall objective as given below:

$$L_{LDMb} := \mathbb{E}_{E(a),b,\varepsilon,t}[\|\varepsilon - \varepsilon_\theta(a_{l_t},t,\mathscr{Z}_\theta(b))\|_2^2] \quad (3)$$

The conditioning mechanism can be represented as:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{h}}\right).V \quad (4)$$

where h denotes the dimension of vector embedding, Q, K, and V are the query, key, and value sequences, respectively. These sequences can further be computed as:

$$Q = W_{Q(i)}.F_i(a_{lt}), K = W_{K(i)}.Z_\vartheta(b), V = W_{V(i)}.Z_\vartheta(b) \quad (5)$$

$F_i(a_{lt})$ signifies the flattened intermediate U-Net representation of the denoising process and $W_Q^{(i)}$, $W_K^{(i)}$ and $W_V^{(i)}$ are the learnable projection matrices with shape $h\times hz$, $h\times hz$, and $h\times h^i_\varepsilon$, respectively.

Text Prompts: In our work, we used text prompts as the conditioning to control the generation process. To construct the prompts, quality information was incorporated in Stage-I prompts whereas pathological and modality-related information was used in Stage-II. The stage-wise prompt construction process is explained below:

Stage-I: *(a) Polyp/non-polyp*: The polyp frames are fed into the model with a prompt incorporating *'with polyp'* substring, whereas non-polyp frames are provided with *'without polyp'* substring. *(b) Good-quality/low-quality*: The polyp frames have additional information about low-quality (frames with artifacts like ghost colors, fecal depositions, low-illumination, etc.) and good-quality (free from artifacts). The text prompts in the former case involve *'low-quality, artifacts'* keywords, whereas *'good-quality, clear'* is part of the latter case.

Stage-II: *(a) Adenomatous/hyperplastic*: To make the model learn and focus on the pathological features, text prompts depicting appropriate condition, *'adenomatous polyp'* or *'hyperplastic polyp'*, were provided. *(b) WLI/NBI*: To generate diverse

Model	Stage	Description
DenseNet-201	I	(i) Used to choose the best iteration of PathoPolyp-Diff training during Stage-I. (ii) Detects polyps, supports t-SNE plot visualization and validates the polyp-characterizing features existing in the generated polyp images. (iii) Results are shown in Table 3 and Fig. 4.
DenseNet-121	II	(i) Used to choose the best iteration of PathoPolyp-Diff training during Stage-II. (ii) Supports t-SNE plot visualization and performs iteration-wise adenomatous and hyperplastic classification with WLI and NBI modalities. (iii) Related results are given in Table 4, Fig. 6 and Fig. 7.
EfficientNet-B0	II	(i) Validates the effectiveness of using synthetic images by classifying adenomatous and hyperplastic classes with WLI and NBI modalities with different proportions of real and synthetic image count in the training set. (ii) Corresponding results are shown in Table 5, Table 6, Fig. 9 and Fig. 10.

Table 1. Description of different models used to validate the outcomes during Stage-I and Stage-II.

polyp images with both WLI and NBI modality, the text prompts at this stage also included *'white light imaging'* or *'narrow band imaging'* substrings.

Cross-Class Label Learning: Cross-Class Label Learning is a two-stage weak-label strategy that integrates (a) knowledge transfer and (b) stochastic regularization mechanism. As a knowledge transfer mechanism, it leverages information learned in Stage-I to enforce generation constraints in Stage-II, where crucial information (image quality) is implicitly needed but explicitly missing. Initially, it forces the text encoder (Z_ϑ) to map the "quality" tokens (*"good-quality, clear"*) into a robust latent embedding space ($y_{Quality}$). In Stage-II, we define the related conditioning vector (y_{CCLL}) as the discrete concatenation of the pathology, modality, and the pre-learned quality tokens. This process embeds the model's generation process within the high-fidelity region of the image manifold.

$$y_{CCLL} = Z_\vartheta(b) = Z_\vartheta(\text{Pathology},\text{Modality},y_{Quality})$$

This transfer acts as a form of weak supervision, where the implicit, missing label ($y_{Quality}$) is injected to constrain the output distribution, ensuring the generation is clinically relevant. Furthermore, this process incorporates a stochastic regularization mechanism, achieved through undersampling (from Stage-I) that ensures the retention of learned features and prevents overfitting to the noise or artifacts present in the limited pathology examples in Stage-II. By retaining a subset of non-polyp and informative polyp images during training, the model is repeatedly exposed to the core structural features learned in Stage-I. This acts as a regularization that prevents catastrophic forgetting of the quality embeddings during fine-tuning. Therefore, it maintains stability while focusing on hard boundary examples.

Training Details

Stage-I: The Stage-I uses a pre-trained Stable Diffusion v1-4[24] model and further fine-tunes it with some text conditions to generate desired colonoscopy images. This stage is focused on developing a model that learns the basic features to differentiate polyp and non-polyp characteristics. In the process, text prompts are used as the conditioning mechanism to control the model output. These text prompts comprise the embeddings pertaining to the strings presented at the first to third levels in Fig. 3, starting from the top. During the fine-tuning of the model, a relatively large-scale dataset is used with polyp/non-polyp classes

wherein the polyps have an additional annotation of low-quality/artifacts (uninformative) and good-quality/clear (informative).

This process allows the model to generate polyp and non-polyp images with specific quality criteria.

Stage-II: In this stage, the pre-trained model of Stage-I is used with the first block locked. The other blocks are further fine-tuned on our desired text conditioning. These conditions are shown in the fourth and fifth levels of Fig. 3. For a successful implementation of cross-class label learning, we undersampled the non-polyp set and informative set of polyps and included them in the training iterations of Stage-II. Presenting these undersampled images allows the model to retain the features pertaining to Stage-I without undergoing overfitting with the new dataset.

The different deep learning models used to validate the outcomes during the above two stages are given in Table 1 with their corresponding objectives. The dataset split details with frame count are summarized in Table 2. More dataset details and the training settings are given in the next section. The source code developed in this work is available at https:

//github.com/Vanshali/PathoPolyp-Diff.

Results

Dataset Details and Training Settings

We used two publicly available datasets, namely, the SUN Database[10,11] and ISIT-UMR Colonoscopy Dataset[9], to evaluate the performance of our proposed model. SUN Database comprises 49,136 polyp frames and 109,554 non-polyp frames. We further used additional annotations of polyp frames for informative/uninformative classes provided by[29] (not provided in the original SUN Database). These additional annotations comprise 31% uninformative polyp frames and are used to train our model in Stage-I. We used the same dataset split as used in[29]. ISIT-UMR Colonoscopy Dataset is an NBI and WLI video

Stage/ Model	Dataset	Videos			Frames			
		Total	Train	Test	Total	Train	Validation	Test
Stage-I ((a) Pathopolyp-Diff, (b) DenseNet-201)	SUN Database	100 cases	-	-	49,136 polyp / 109,554 non-polyp (random split)	80%	10%	10% (For (b), synthetic images are tested.)
Stage-I ((a) Pathopolyp-Diff, (b) FFT (only Test split applicable))	SUN Database	100 cases	-	-	49,136 polyps (31% uninformative frames, remaining are informative frames, random split)	80%	10%	10 %
Stage-II (PathoPolyp-Diff)	ISIT-UMR	40 AD / 21 HP (for each WLI and NBI)	29 AD / 15 HP	-	-	12,883 AD / 12,987 HP (both WLI and NBI)	-	-
Stage-II ((a) DenseNet-121, (b) EfficientNet-B0)	ISIT-UMR	40 AD / 21 HP (for each WLI and NBI)	15 AD / 15 HP	6 AD / 6 HP	-	Variable number of frames from each video (16, 23 and 64). More details in Table 5.	-	4,376 AD / 2,597 HP (for each WLI and NBI), For (a), synthetic images are tested.

Table 2. Dataset details with video-wise and frame-wise split for different models. AD and HP refer to adenoma and hyperplastic, respectively.

dataset and consists of labels for hyperplastic, adenoma and sessile classes. We used the first two of these classes to train our model in Stage-II. This dataset was split into separate, non-overlapping training and test sets to ensure that no frames or

videos from the training set were used for testing. The same split was used during pathology-based classification. We converted 40 adenoma and 21 hyperplastic video streams of each NBI and WLI modalities into frames. We performed a train and test split (video level split) for each class for both modalities, thus creating a split of ratio 29:11 and 15:6 for adenoma and hyperplastic classes, respectively. 29 adenoma and 15 hyperplastic videos of both modalities are used for training PathoPolyp-Diff. The ISIT-UMR dataset has a notable class imbalance between adenoma and hyperplastic classes. These are pathology-based classes, and differentiating them is a complex task, which could unnecessarily impact the training and validation of our proposed diffusion model. Therefore, we undersampled video frames of adenoma class during PathoPolyp-Diff training. Further, during the classification task, we selected an equal number of variable frames (16, 32, or 64) from each video, therefore, to avoid class imbalance, an equal number of videos of both classes were randomly selected. 15 videos (a subset of the previous 29 adenomas) and 6 videos of each pathology and modality were randomly selected to form the train set and test set, respectively. Note that the test set is mutually exclusive to the training sets of both the diffusion model and the classifier. Due to the lack of existing text-to-image generative frameworks specifically designed for pathological colonoscopy synthesis, a direct fair comparison with prior work was not feasible. In the absence of such directly comparable architectures, we established a standard benchmark,

where the baseline represents the conventional training approach using only real images.

During PathoPolyp-Diff training in Stage-I, a pre-trained Stable Diffusion v1-4 model was loaded, which was further fine-tuned on the SUN Database with a learning rate, batch size and resolution set to $1e^{-06}$, 15, and 512, respectively. A lower learning rate allows stable training as mentioned in the literature[24] and a resolution of 512 maintains compatibility with the resolution on which the Stable Diffusion v1-4 model was originally trained. The best model chosen for subsequent fine-tuning in Stage-II was identified after 8,000 iterations (based on t-SNE plots and Table 3). In Stage-II, ISIT-UMR Colonoscopy Dataset is used for training with the same learning rate and one block of U-Net locked. The training of EfficientNet-B0 classifier involves 50 epochs and 0.00005 learning rate. All the implementations are carried out using the PyTorch framework and experiments are executed on NVIDIA A100 and NVIDIA Titan-Xp GPU for PathoPolyp-Diff and classification tasks, respectively.

Evaluation Metrics: In this work, we adopted some standard metrics used for classification. It includes precision, recall, F1-score, and balanced accuracy. The first three are commonly used to evaluate any classification model. The last metric, i.e. balanced accuracy, is generally used when an imbalance is encountered in data distribution, and the objective focuses on both minority and majority classes with equal importance during evaluation. This scenario aligns with our case, where we consider both adenomatous and hyperplastic classes to be equally important, as the clinical treatments depend on the diagnosed pathology class. The balanced accuracy can be defined as an arithmetic mean of specificity (true negative rate) and sensitivity

(true positive rate). In addition, we used Kernel Inception Distance (KID)[30] for quality assessment of the generated images. It quantifies the dissimilarity between the generated and real data distributions. Unlike the standard KID metric, which uses Inception-v3 as the feature extractor, we used the DenseNet model inspired by the literature[29,31], where it achieved remarkable performance in a similar domain.

Model Performance

Stage-I: To evaluate the performance of our model in generating polyp and non-polyp images and to select the best model for the subsequent training process, we used four assessment metrics, namely, KID, precision, recall, and F1-score. For KID computation, we processed both real and generated images through the DenseNet model[29,31] as discussed in the above section

Prompt	Metrics	Behavior	Iterations									
			1K	2K	3K	4K	5K	6K	7K	8K	9K	10 K
colonoscopy image with polyp, good-quality, clear	Precision	↑	0.8671	0.9085	0.9617	0.9720	0.9850	0.9858	0.9961	0.9859	0.9761	0.9899
	Recall	↑	0.9567	0.9933	0.8367	0.9267	0.8733	0.9267	0.8467	0.9300	0.9533	0.9800
	F1-score	↑	0.9097	0.9490	0.8948	0.9488	0.9258	0.9553	0.9153	0.9571	0.9646	0.9849
	KID	↓	0.045	0.059	0.055	0.075	0.053	0.071	0.048	0.051	0.049	0.064
	Intra-class variation	↑	0.5907	0.5964	0.6047	0.6091	0.6166	0.6097	0.6029	0.5924	0.5918	0.5866
colonoscopy image without polyp	KID	↓	0.027	0.041	0.037	0.034	0.067	0.051	0.051	0.045	0.055	0.057
	Intra-class variation	↑	0.6314	0.6431	0.6516	0.6516	0.6539	0.6745	0.6492	0.6676	0.6671	0.6698

	Average KID	↓	0.0360	0.0500	0.0460	0.0545	0.0600	0.0610	0.0495	0.0480	0.0520	0.0605

Table 3. Iteration-wise quality assessment of generated images in Stage-I. This assessment is done using KID (similarity with real images), precision, recall, F1-score (polyp/non-polyp characterizing features) and intra-class variation using LPIPS. ↓ and ↑ denote 'lower is best' and 'higher is best', respectively. Values in bold denote the best outcomes.

to obtain their respective feature vectors. The KID score was then calculated using the squared maximum mean discrepancy[32] between the feature distributions of the real and synthetic image feature vectors. A lower KID score indicates that the generated images closely resemble real data in terms of feature distributions, indicating higher quality and diversity. Initially, we trained the model for 10,000 iterations and selected the model after every 1,000 iterations for testing. The corresponding assessment results are given in Table 3. As our main focus is to obtain good-quality polyp images, we generated and validated images with substring *"good-quality, clear"*. Additionally, we used a DenseNet-201[33], pre-trained with the same training settings as used by[29] to detect polyps. This model validates the polyp-characterizing features existing in the generated polyp images. For this,

we tested generated polyp and non-polyp images using DenseNet-201. It can be observed that the lowest (also the best) average KID of 0.036 is obtained at 1,000th iteration; however, it reported a low F1-score. Similar outcomes are achieved with the next lowest KID. Contrarily, the highest F1-score in the 10,000th iteration is obtained with 0.0605 KID, the second least favorable among all iterations. Note that while we report intra-class variations using LPIPS[34], our evaluation primarily prioritizes KID and F1-score, as our main objective is to generate clinically relevant and diagnostically useful images.

To study the reasoning behind the contradictory results, we plotted t-SNE embeddings, shown in Fig. 4. It can be observed that in the initial iterations (1,000 to 5,000), the polyp and non-polyp features of synthetic data are not entirely distinct; therefore, the associated F1-scores are relatively low. At the same time, they are finely overlapping with their real counterparts, therefore, resulting in lower KID. To establish a trade-off between the KID and F1-score, we leveraged the visualization capabilities of t-SNE plots. Finally, the model at 8,000th iteration is selected as its KID score (in terms of all categories, i.e., polyp, non-polyp and average) is higher than the average score computed over each corresponding KID category. For instance, the average KID_{polyp} is calculated as $KID_{polyp1k}$ + $KID_{polyp2k}$ +$KID_{polyp3k}$…+$KID_{polyp10k}$, which comes out to be 0.57. It can be noted that $KID_{polyp8k} < KID_{polyp}$, and a similar observation can be identified in other categories. Moreover, the t-SNE plots signify that after 8,000th iteration, the synthetic polyp and non-polyp features start to deviate from the feature space, representing their real counterparts. As one class of synthetic features is still far apart from the other class of synthetic features, the F1-score is higher at the last iterations. The above analysis establishes a relationship between the F1-score and the KID score. While the F1-score effectively measures the distinction between the two classes, the KID score assesses their similarity to real counterparts. Notably, these metrics may conflict in scenarios where the classes exhibit a fair overlap with the real counterparts yet remain poorly separated, or when they are well-separated but lack alignment with real data distributions. To further support these results, we included some sample images from the two extreme cases, i.e. iterations 1,000th and 10,000th, and the best iteration 8,000th in Fig. 5. In the initial iteration, the model demonstrates an understanding of the fundamental structure; however, it lacks refinement in texture and overall appearance. Similarly, in the last iteration, the background structure appears distorted, exhibiting unusual color patterns. The samples corresponding to 8,000th iteration show visually appealing outcomes, retaining texture and structural details. Overall, it can be observed from the KID scores that the images generated by PathoPolyp-Diff were comparable to real images, suggesting that our model effectively captures the underlying patterns and variations present in the colonoscopy images.

Impact of Negative Prompt: Negative prompt is an additional parameter which guides the image generation process *not* to include some specific objects or characteristics. It can assist in eliminating unwanted elements from the synthetic images. Hence, to further improve the quality of generated images, we evaluated our model on various negative prompts, including

"low-quality", *"blur"* and *"low-quality, blur"*. This approach is experimented with both polyp and non-polyp frames, and quality assessment is performed using the fast Fourier transform (FFT). This metric quantifies the blurriness content of the given image. The results shown in Fig. 8 present a significant improvement when a normal text prompt is combined with our specific quality-based negative prompt. The best combination is achieved when we club text prompt with *"blur, low-quality"* negative prompt. Therefore, the subsequent experiments include this specific negative prompt to enhance quality, as it helps the

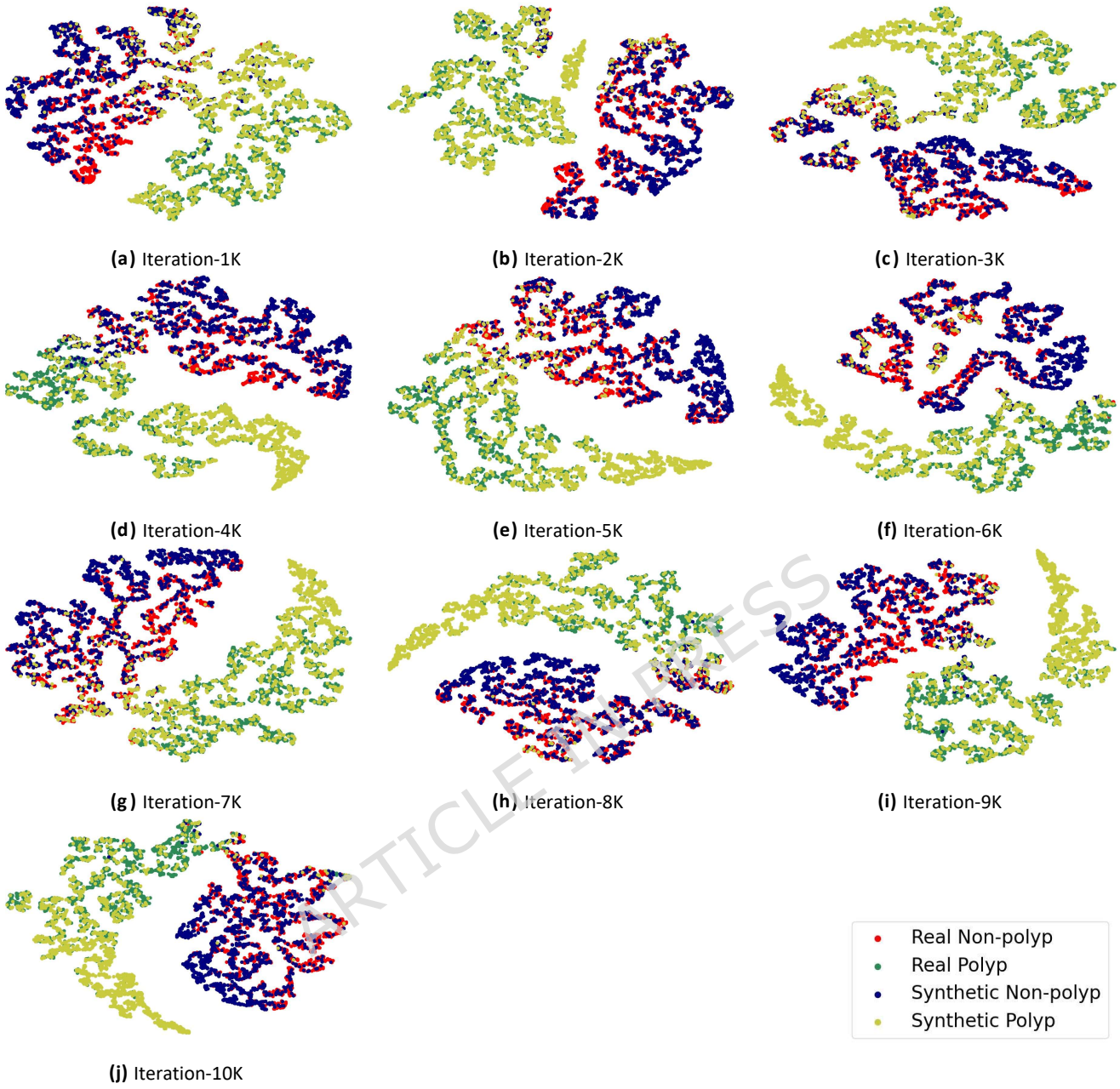
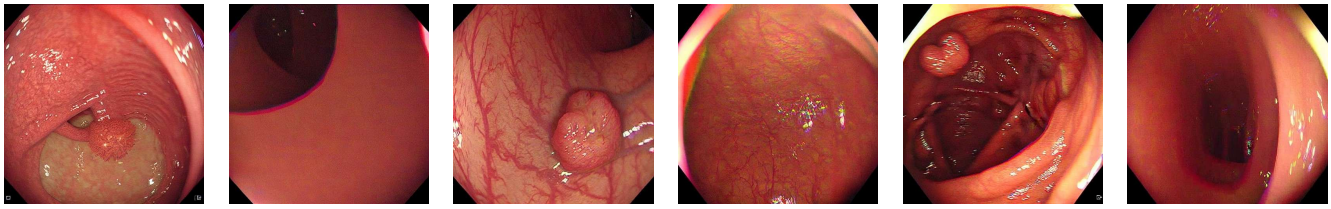


Figure 4. Iteration-wise two-dimensional t-SNE embeddings to visualize the data points pertaining to synthetic and real polyp/non-polyp images. Iteration-8K is selected for further experiments.

model avoid generating blurred or low-quality images.

Stage-II: To assess the quality of images generated in Stage-II, specifically in terms of pathological characteristics, we performed two different binary classifications. The corresponding purposes of these classifications are (a) To select the best model among different iterations (validation similar to Stage-I) and (b) To verify how the synthetic images impact the classification performance if used for augmentation. In the first case, we used a pre-trained DenseNet-121 model intended to classify adenomatous and hyperplastic classes with WLI and NBI modalities. The pre-training is conducted on the same training dataset that is later used to train the EfficientNet-B0 classifier, however, with all videos and frames. Also, note that the training data for the diffusion model and the classifier overlap, while the classifier's test set remains entirely separate from the training data of both models. This overlap in the training dataset does not affect the evaluation, as the synthetic

images generated by the diffusion model are used solely to augment the classifier's training data, ensuring fair evaluation without bias. The synthetic images generated from every 1000th iteration are evaluated using this model and the related results are shown in



(a) (b) (c) (d) (e) (f)

Figure 5. Sample generated images from iterations (a) 1K polyp, (b) 1K non-polyp, (c) 8K polyp, (d) 8K non-polyp, (e) 10K polyp, and (f) 10K non-polyp.

Class	Metrics	Behavior	Iterations									
			1K	2K	3K	4K	5K	6K	7K	8K	9K	10 K
Adenoma	Precision	↑	0.50	0.54	0.57	0.65	0.69	0.71	0.65	0.59	0.61	0.58
	Recall	↑	0.99	0.98	0.99	0.99	0.98	0.97	0.98	0.99	0.99	0.99
	F1-score	↑	0.67	0.69	0.73	0.79	0.81	0.82	0.78	0.74	0.76	0.73
Hyperplastic	Precision	↑	0.75	0.90	0.98	0.97	0.97	0.96	0.95	0.96	0.98	0.95
	Recall	↑	0.02	0.15	0.26	0.48	0.55	0.61	0.48	0.33	0.37	0.29
	F1-score	↑	0.04	0.26	0.42	0.64	0.70	0.74	0.64	0.49	0.54	0.45

Table 4. Class-wise quality assessment of generated images after every 1000 iterations during Stage-II. ↑ denotes 'higher is best'. Values in bold denote the best outcomes.

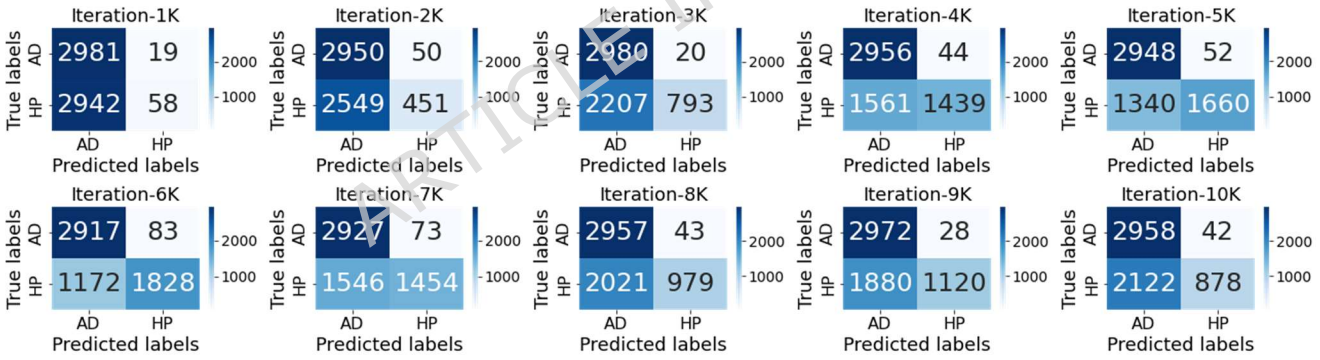


Figure 6. Confusion matrices to validate the iteration-wise performance of our model in generating adenomatous/hyperplastic polyp images with NBI/WLI imaging modalities. AD and HP refer to adenoma and hyperplastic, respectively.

Table 4. Similar to Stage-I, we plotted t-SNE feature embeddings to analyze the overlay and proximity of synthetic features to real images. As illustrated in Fig. 7, the pathology-specific features are not learnt in the initial epochs. Hence, a significant overlap is observed among the generated images from different classes. It can also be inferred from the findings presented in Table 4, where DenseNet-121 exhibited significant challenges in effectively distinguishing between the two classes until 3000th iteration. These outcomes are further supported by the confusion matrices in Fig. 6. These matrices demonstrate the biased shift of the model towards the adenomatous class, which gradually improves with increasing epochs and, after some epochs, again shows the same biased performance. This trend is observed due to the deviation of synthetic data from the expected pathological behavior as we train the model after a certain number of iterations. This analysis is based on the last few plots in Fig. 7. Considering the results in Table 4, Fig. 6, and Fig. 7, we selected the model at 6000th iteration which reports the highest F1-score for both the classes (adenoma: 0.82, hyperplastic: 0.74).

In the second case (i.e. case (b)), we used the synthetic images to augment the real data (using ISIT-UMR dataset). The effectiveness of using synthetic images is validated using a binary classifier, EfficientNet-B0[35]. In addition to validating the synthetic data inclusion, we also compared the quality of synthetic images obtained using two different text prompts (with and without cross-class label learning). Table 5 shows the associated frame-wise results with various data proportions. Starting with real data samples of 16, 32 and 64 frames per video, we subsequently increased the sample count by adding synthetic images in *ix* proportion, where *i* = {1,2,3} and *x* is the real frame count. This procedure is followed for A and B text prompts, which can be defined as *"colonoscopy image with 'y' polyp, 'z'"* and *"colonoscopy image with 'y' polyp, 'z', good-quality, clear"*,

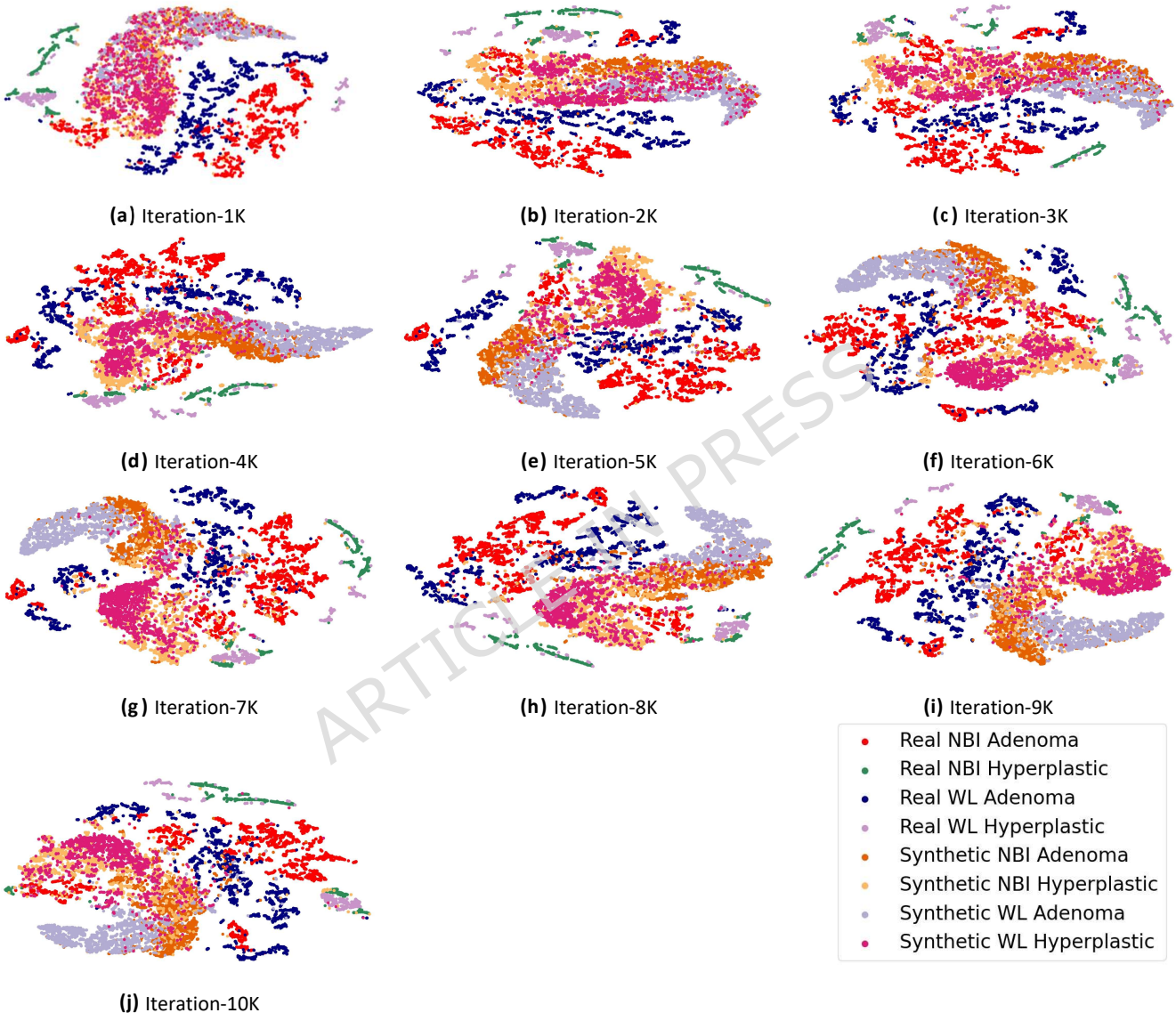


Figure 7. Iteration-wise two-dimensional t-SNE embeddings to visualize the data points pertaining to synthetic and real adenomatous/hyperplastic images involving NBI/WLI imaging modalities. Iteration-6K is selected since the two classes are best differentiable here (also with high F1-score).

respectively where 'y' denote adenomatous/hyperplastic and 'z' denote NBI/WLI. It is noteworthy that the label *"good-quality, clear"* used in the text prompt B is not directly related to the training samples provided to PathoPolyp-Diff during Stage-II; instead, they are learnt from a different dataset through cross-class label learning during Stage-I. The comparative analysis of synthetic images pertaining to the two text prompts aims to assess the effectiveness of cross-class label learning. As shown in Table 5, with 16 real images per video, adding an equal number of synthetic images improves the

frame-wise balanced accuracy by 1.14% to 4.75% with NBI and WLI. A similar increase of 2.66% to 3.93% is reported when the ratio of synthetic data is doubled. With a further increase, i.e., when the proportion of synthetic data is three times, the results are enhanced by up to 7.91%. However, with 32 or 64 real images per video, the performance increase is relatively less significant and limited to about 3.5%. Therefore, it can be inferred that a relatively significant performance gain is achieved when a substantially small real dataset is merged with synthetic data. Further increasing the real dataset shows comparatively less improvement (without a monotonic trend) with a similar data augmentation approach. The trend shown in Table 5 also signifies that over-training the model with synthetic images degrades or saturates the results after a certain proportion, which aligns with the investigation conducted in literature in a similar domain[36,37].

Impact of Cross-class Label Learning: We further performed validation and analysis for the cross-label learning approach.

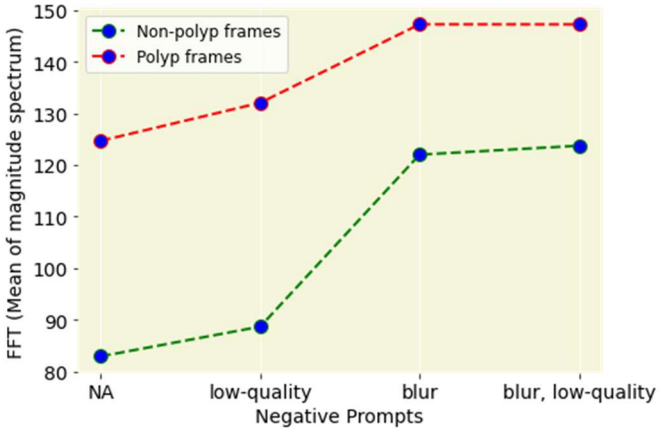


Figure 8. Quality assessment for validating the impact of negative prompt using FFT.

It can be observed that in most cases, the balanced accuracy and class-wise F1-scores using text prompt B (cross-class label learning) are comparatively higher than that of text prompt A (without cross-class label learning). Some notable improvements in balanced accuracy include 0.6192 to 0.6614 (difference of +4.22%, 95% CI: 1.5%, 6.91% , p-value = 0.0068) using 64 real samples per video with an equal proportion of synthetic data, 0.5561 to 0.6254 (difference of +6.93%, 95% CI: 5.01%, 8.85 %, p-value < 0.0001) using 16 real samples per video with two times samples of synthetic data, and 0.6228 to 0.6983 (difference of +7.55%, 95% CI: 2.88%, 12.22%, p-value = 0.0058) using 16 real samples per video with three times samples of synthetic data. These outcomes signify that the quality of generated data can be improved with variations in the input text prompts. The consistent superiority of prompt B serves as an ablation analysis of our two-stage distillation. By incorporating "goodquality, clear" tokens learned in Stage-I, guiding the diffusion process toward clinically relevant features while suppressing uninformative artifacts. Moreover, the labels used in these text prompts can be indirectly inferred from other classes, thus reducing the requirements of annotated data for each scenario.

Video-wise Analysis with Statistical Significance Test

A comparative study has been conducted on patient-wise results. Although the training experiments are performed on the frame level, the inference is computed on both the frame and video levels. A majority voting scheme has been adopted for such computations. For instance, a video is labelled as the class 'adenomatous' if the majority of frames (> mean number of total video frames) are predicted as adenomatous. These results are provided in the Table 6a and Table 6b. Additionally, a statistical significance test is conducted using a two-tailed t-test, which indicates the significance of an increase or decrease in video-wise outcomes and reports if this change is insignificant. This test calculates the p-values between all possible combinations of the data proportions used in our work. The associated p-values are depicted in Fig. 9a and Fig. 10a. Using the WLI modality, we observed that the best outcomes are obtained using 32 real images per video and 16 or 32 real images per video combined with synthetic images generated from text prompt B in equal proportion. The last two cases align with the second-best results in frame-level evaluation presented in Table 5. Moreover, considering the case with 16 real images per video and an equal proportion of synthetic samples, video-level analysis reports a significant improvement of 15% (p-value = 0.037). Similar to frame-level analysis, the video-level validations signify that the performance improvement with synthetic data is reduced by increasing the real sample count. Although with 64 real images per video, an increase of 3.34% is observed, this increase is not statistically significant (p-value = 0.587). We further examined the performance difference between the two text prompts, A and B, to evaluate the video-level cross-class label learning ability. Our analysis demonstrates that synthetic images generated using text prompt B are superior and statistically significant to those generated using text prompt A. This observation is supported by some of the notable improvements that include 0.6 to 0.7833 (difference of +18.33%, 95% CI: 3.43%, 33.23%, p-value = 0.022) using 16 real samples per video with an equal number of synthetic samples, 0.6 to 0.7167 (difference of +11.67%, 95 %

CI: 1.09%, 22.25%, p-value = 0.034) using 64 real samples per video with an equal number of synthetic samples, and 0.6333 to 0.7167 (difference of +8.34%, 95% CI: 1.63%, 15.05%, p-value = 0.020) using 32 real samples per video combined with three times as many synthetic samples.

Using the NBI modality, the overall performance trend with data proportion is similar. Also, the comparison between the real data and a combination of real and synthetic data shows a similar trend as observed using the WLI modality. However, in the NBI modality, the synthetic data generated using text prompt A presented better outcomes than text prompt B in some cases. Although the best results are obtained using text prompt A, the average performance over all data proportions is

Modality	(Real + Synthetic)		Text Prompt	Precision	Recall	F1-score	Precision	Recall	F1-score	Balanced Accuracy
NBI	x=16 images per video	x	- (baseline)	0.8173±0.026	0.2897±0.035	0.4266±0.037	0.4757±0.009	0.9079±0.020	0.6242±0.007	0.5988±0.012
		x+x	A	0.7876±0.122	0.3194±0.111	0.4360±0.099	0.4689±0.006	0.8449±0.121	0.6005±0.026	0.5821±0.007
		x+2x	B	0.8146±0.080	0.3295±0.050	0.4671±0.056	0.4853±0.023	0.8910±0.059	0.6281±0.031	0.6102±0.037
		x+3x	A	0.6596±0.020	0.4131±0.025	0.5077±0.021	0.4580±0.012	0.6991±0.029	0.5533±0.016	0.5561±0.015
			B	0.8299±0.042	0.3569±0.032	0.4977±0.027	0.4965±0.007	0.8939±0.040	0.6382±0.012	0.6254±0.011
			A	0.7359±0.059	0.3529±0.055	0.4730±0.046	0.4712±0.014	0.8133±0.076	0.5959±0.028	0.5831±0.022
			B	0.8133±0.113	0.3524±0.047	0.4857±0.027	0.4867±0.010	0.8680±0.097	0.6227±0.033	0.6102±0.025
	x=32 images per video	x	- (baseline)	0.8099±0.083	0.3320±0.036	0.4690±0.040	0.4840±0.020	0.8843±0.068	0.6253±0.032	0.6082±0.034
		x+x	A	0.7713±0.078	0.3970±0.073	0.5170±0.051	0.4909±0.007	0.8192±0.093	0.6124±0.025	0.6081±0.014
		x+2x	B	0.7634±0.055	0.3980±0.015	0.5222±0.008	0.4917±0.011	0.8221±0.054	0.6151±0.023	0.6101±0.021
		x+3x	A	0.7562±0.029	0.4155±0.024	0.5358±0.022	0.4958±0.012	0.8101±0.033	0.6150±0.017	0.6128±0.017
			B	0.8240±0.066	0.4160±0.041	0.5510±0.035	0.5137±0.018	0.8700±0.064	0.6455±0.028	0.6430±0.029
			A	0.7797±0.083	0.3841±0.036	0.5115±0.024	0.4905±0.013	0.8371±0.074	0.6180±0.028	0.6106±0.024
			B	0.8464±0.090	0.3805±0.043	0.5221±0.040	0.5060±0.020	0.8948±0.068	0.6460±0.030	0.6377±0.031
	x=64 images per video	x	- (baseline)	0.8889±0.035	0.3698±0.022	0.5216±0.019	0.5124±0.005	0.9338±0.025	0.6615±0.007	0.6516±0.008
		x+x	A	0.7704±0.026	0.4129±0.032	0.5368±0.027	0.4995±0.011	0.8255±0.033	0.6221±0.014	0.6192±0.014
		x+2x	B	0.8463±0.062	0.4405±0.021	0.5780±0.011	0.5277±0.010	0.8523±0.060	0.6601±0.024	0.6614±0.022
		x+3x	A	0.7964±0.031	0.4120±0.016	0.5426±0.011	0.5062±0.007	0.8501±0.034	0.6344±0.014	0.6310±0.013
			B	0.7818±0.056	0.4033±0.013	0.5314±0.014	0.4988±0.013	0.8381±0.051	0.6252±0.024	0.6207±0.023
			A	0.8348±0.080	0.4086±0.012	0.5477±0.023	0.5132±0.021	0.8810±0.069	0.6483±0.035	0.6448±0.036
			B	0.7900±0.078	0.3922±0.028	0.5218±0.014	0.4959±0.013	0.8447±0.073	0.6245±0.029	0.6185±0.025
WLI	x=16 images per video	x	- (baseline)	0.7856±0.064	0.5689±0.084	0.6583±0.072	0.4237±0.068	0.6695±0.096	0.5176±0.074	0.6192±0.073
		x+x	A	0.7869±0.057	0.6247±0.058	0.6937±0.031	0.4373±0.041	0.6268±0.129	0.5126±0.062	0.6257±0.051
		x+2x	B	0.8563±0.039	0.5296±0.051	0.6521±0.031	0.4444±0.010	0.8037±0.070	0.5714±0.015	0.6667±0.013
		x+3x	A	0.7709±0.037	0.7551±0.049	0.7616±0.028	0.4944±0.054	0.5144±0.111	0.5013±0.073	0.6348±0.048
			B	0.8385±0.044	0.5635±0.113	0.6658±0.072	0.4504±0.028	0.7534±0.119	0.5593±0.026	0.6585±0.020
			A	0.7652±0.017	0.7196±0.052	0.7407±0.028	0.4687±0.029	0.5260±0.066	0.4937±0.035	0.6228±0.023
			B	0.8621±0.053	0.6302±0.156	0.7154±0.099	0.5079±0.079	0.7665±0.130	0.6010±0.040	0.6983±0.039*
	x=32 images per video	x	- (baseline)	0.8460±0.033	0.6760±0.089	0.7475±0.046	0.5188±0.047	0.7291±0.089	0.6018±0.019	0.7026±0.016
		x+x	A	0.7983±0.030	0.7695±0.089	0.7801±0.034	0.5493±0.050	0.5751±0.126	0.5514±0.046	0.6723±0.024
		x+2x	B	0.8328±0.034	0.7009±0.023	0.7605±0.020	0.5199±0.017	0.6952±0.080	0.5939±0.039	0.6981±0.032
		x+3x	A	0.7868±0.024	0.7011±0.093	0.7395±0.059	0.4920±0.075	0.5957±0.051	0.5351±0.047	0.6484±0.044
			B	0.8454±0.040	0.5983±0.092	0.6963±0.057	0.4728±0.037	0.7591±0.097	0.5793±0.032	0.6787±0.032
			A	0.7960±0.015	0.7799±0.089	0.7857±0.045	0.5628±0.079	0.5714±0.068	0.5602±0.029	0.6756±0.025
			B	0.8542±0.056	0.5664±0.186	0.6626±0.113	0.4701±0.072	0.7675±0.157	0.5696±0.020	0.6670±0.023
	x=64 images per video	x	- (baseline)	0.8007±0.053	0.5690±0.149	0.6574±0.120	0.4432±0.084	0.7023±0.095	0.5389±0.073	0.6357±0.075
		x+x	A	0.7657±0.035	0.7225±0.066	0.7408±0.019	0.4634±0.018	0.5171±0.139	0.4824±0.073	0.6198±0.038
		x+2x	B	0.8445±0.029	0.4436±0.062	0.5792±0.052	0.4096±0.019	0.8229±0.050	0.5463±0.019	0.6333±0.022
		x+3x	A	0.7767±0.025	0.6606±0.100	0.7107±0.060	0.4570±0.054	0.5928±0.083	0.5112±0.037	0.6267±0.035
			B	0.8183±0.062	0.4888±0.120	0.6016±0.086	0.4086±0.027	0.7499±0.143	0.5247±0.042	0.6194±0.036
			A	0.7559±0.011	0.6762±0.023	0.7135±0.012	0.4343±0.013	0.5320±0.038	0.4778±0.020	0.6041±0.013
			B	0.8230±0.043	0.5547±0.075	0.6589±0.044	0.4362±0.021	0.7361±0.113	0.5451±0.042	0.6454±0.033

Imaging Training sample count Adenoma Hyperplastic

Table 5. Classification results using different proportions of real and synthetic images. Text prompt A and B stand for *"colonoscopy image with 'y' polyp, 'z'"* and *"colonoscopy image with 'y' polyp, 'z', good-quality, clear"*, respectively where 'y' denote adenomatous/hyperplastic and 'z' denote NBI/WLI. The values represent mean ± standard deviation. In 'x+ix', the first term denotes the real frame count and the second term represents the synthetic image count. i is the ratio in which synthetic images are added. Values in bold and underlined denote the best and second-best outcomes, respectively. * indicates the maximum performance gain relative to the baseline (7.91%). same for both text prompts. Moreover, it is noteworthy that the difference is not statistically significant in almost every case. Some such examples include the improvement from 0.7167 to 0.75 (difference of +3.33%, 95% CI: -8.2028% to 14.86 %, p-value = 0.524) using 16 real images per video and an equal number of synthetic images, 0.7833 to 0.80 (difference of +1.67 %, 95% CI: -5.04%, 8.38%, p-value = 0.580) using 32 real images per video and twice as many synthetic images, and 0.7667 to 0.8167 (difference of +5%, 95% CI: -0.40%, 10.40%, p-value = 0.067) using 64 real images per video and twice/thrice as many synthetic images. Despite such change in trend in video-wise analysis, it can be observed from Table 5 that on a frame-wise level, most of the cases favored text prompt B over text prompt A. This inconsistent shift can be due to the fact that control over diffusion models is limited and also depends on the seed value. The original dataset used in Stage-I with annotations based on the quality (good-quality, clear/ low-quality) comprises WLI images, whereas the WLI and NBI images used in Stage-II lack such annotations. It was a relatively simple task for the model to generate a combination of WLI and good-quality data. On the contrary, the constrained control over the generated data occasionally resulted in blending WLI characteristics into some NBI images. This inconsistency emerged from the association between WLI and quality learned during Stage-I training. Random seed initialization and limited control over diffusion models resulted in some arbitrary outcomes with text prompt B in the case of NBI images. This justification is supported by the qualitative outcomes (shown in Fig. 11q-11r) discussed in the next section.

(Real + Synthetic)	Text Prompt	Balanced Accuracy
x	-	0.6333±0.126
x+x	A	0.6±0.137
x+2x	B	0.7833±0.046*
x+3x	A	0.6167±0.046
	B	0.7167±0.046
x=16 images	A	0.6±0.109
per video	B	0.75±0.083
x	-	0.7833±0.046
x+x	A	0.7167±0.119
x+2x	B	0.7833±0.047
x+3x	A	0.6333±0.095
	B	0.7333±0.037
x=32 images	A	0.6333±0.046
per video	B	0.7167±0.046
x	-	0.6833±0.109
x+x	A	0.6±0.070
x+2x	B	0.7167±0.075
x+3x	A	0.5667±0.037
	B	0.6167±0.095
x=64 images	A	0.5333±0.046
per video	B	0.6333±0.095

Training sample count

(Real + Synthetic)	Text Prompt	Balanced Accuracy
x	-	0.6667
x+x	A	0.7167±0.095
x+2x	B	0.75±0.059
x+3x	A	0.6833±0.037
	B	0.7667±0.037
x=16 images	A	0.7833±0.046
per video	B	0.75
x	-	0.7333±0.070
x+x	A	0.7333±0.070
x+2x	B	0.7667±0.037
x+3x	A	0.8±0.046
	B	0.7833±0.046
x=32 images	A	0.7667±0.037
per video	B	0.7833±0.046
x	-	0.8167±0.037
x+x	A	0.8167±0.037
x+2x	B	0.8±0.046
x+3x	A	0.8167±0.037
	B	0.7667±0.037
x=64 images	A	0.8167±0.037
per video	B	0.7667±0.037

Training sample count

(a) **(b)**

Table 6. Video-wise results using (a) WLI modality and (b) NBI modality. Values in bold denote the best outcomes. * indicates the maximum performance gain using prompt B (cross-class label learning) compared to prompt A (18.33%, p-value = 0.022).

Node 1	Prompt A: "colonoscopy image with adenomatous polyp, narrow band imaging"	(m)-(n)
Node 2	Prompt B: "colonoscopy image with adenomatous polyp, narrow band imaging, good-quality, clear"	(e)-(f)
Node 3	Prompt A: "colonoscopy image with adenomatous polyp, white light imaging"	(i)-(j)
Node 4	Prompt B: "colonoscopy image with adenomatous polyp, white light imaging, good-quality, clear"	(a)-(b)
Node 5	Prompt A: "colonoscopy image with hyperplastic polyp, white light imaging"	(k)-(l)
Node 6	Prompt B: "colonoscopy image with hyperplastic polyp, white light imaging, good-quality, clear"	(c)-(d)
Node 7	Prompt A: "colonoscopy image with hyperplastic polyp, narrow band imaging"	(o)-(p)
Node 8	Prompt B: "colonoscopy image with hyperplastic polyp, narrow band imaging, good-quality, clear"	(g)-(h)
Token Combination (Ref: Fig 3)	Text Prompt Strategy	Visual Attribute Mapping (Ref: Fig 11)

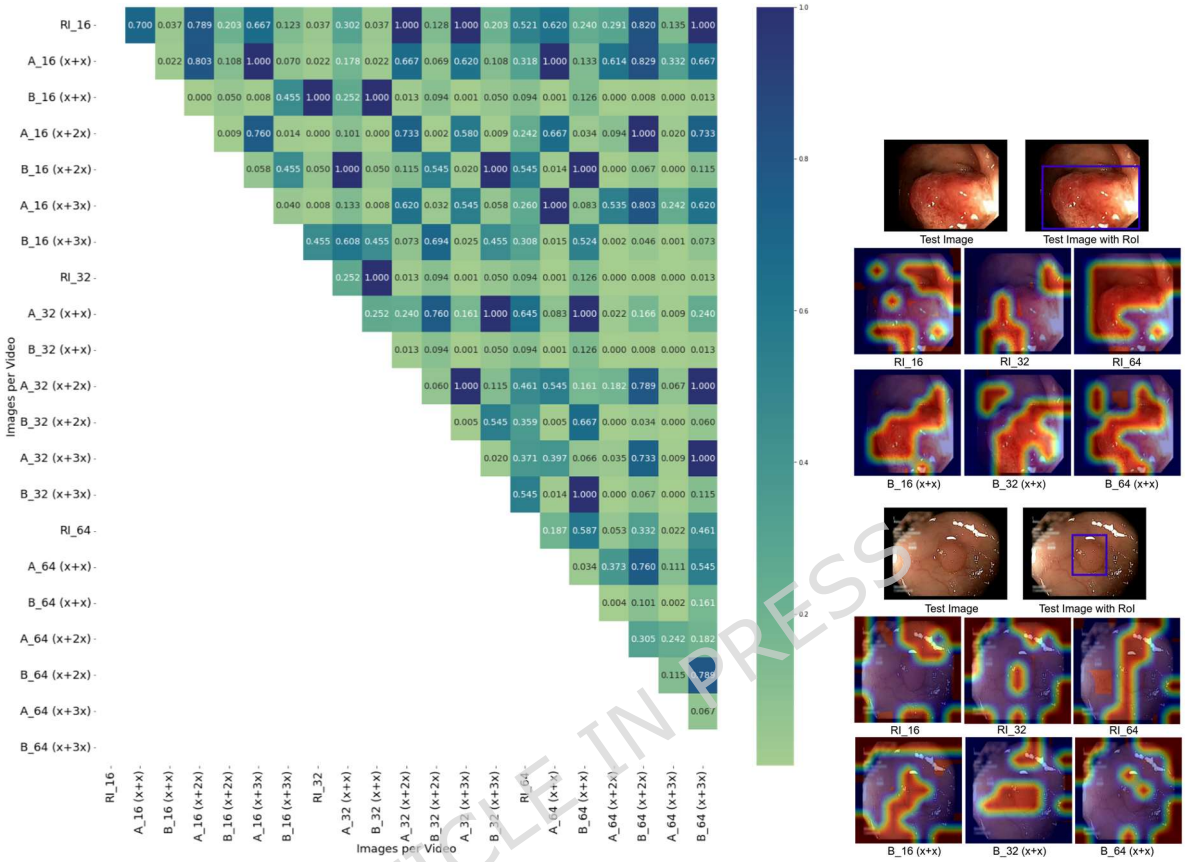
Table 7. Mapping of prompt configurations to synthetic output nodes. This table defines the correspondence between the semantic text prompts (Figure. 3) and the generated images shown in Figure 11. Node indices (1–8) represent the specific feature transmission paths; odd nodes denote standard prompts (Prompt A), while even nodes denote the proposed Cross-Class Label Learning prompts (Prompt B), which incorporate quality-refinement tokens ("good-quality, clear") distilled from Stage-I.

Qualitative Results and Interpretability through Visualization

In addition to the quantitative analysis, we examined the qualitative results and further studied the related heatmaps for visualization and interpretability. The heatmaps were obtained using GradCAM on EfficientNet-B0. After selecting a target layer, heatmaps were generated by attributing gradients to that layer, highlighting the most influential regions in the input images. The outcomes were then overlaid on the original images. The heatmaps pertaining to the scenarios involving real images and also those related to some cases involving synthetic images are shown in Fig. 9b and Fig. 10b. These heatmaps illustrate the region the classification model focuses on before providing the final prediction scores. It can be observed that the classifier learns the complex polyp-specific features better when trained using a more diverse set of polyp images obtained using PathoPolyp-Diff. However, the classifier's performance drops in identifying polyp features when the count of synthetic images increases. This decline in performance could be because synthetic images might carry noise and can not exactly replicate real image characteristics; thus, added noise could deviate the model after a certain limit.

Further, we analyzed the generated images for qualitative analysis. It can be observed that the synthetic images obtained using text prompt B are visually more appealing than those produced with text prompt A. The texture is more prominent and clear in images shown in Fig. 11a to Fig. 11h compared to those presented in Fig. 11i to Fig. 11p. Moreover, in both scenarios, qualitatively, the generated images are close to real images in terms of structure, color and texture. The fundamental color criteria that differentiate NBI and WLI remain consistently evident in the images, making them easily distinguishable. These generated samples can be easily mapped to the nodes shown in Figure. 3 and exact prompts using Table 7. However, as already discussed, text prompt B with NBI images fails for some samples, as can be inferred from Fig. 11q and Fig. 11r. This failure is attributed to the limited control over the image generation and dependency on seed

initialization. Note that the pathology-focused data used in Stage-II has both WLI and NBI images, but none of the images have annotations based on quality. Still, the model learnt the relation between quality and WLI easily because of the relation developed earlier in Stage-I. Therefore, in Stage-II, our model more precisely established the relation between WLI, quality, and pathology. Due to



(a) p-value obtained using two-tailed t-test for statistical significance analysis. The values are rounded off to 3 decimal places. The label names used in rows and columns can be read as "Text Prompt_Sample Count per Video (Real Image Count + Synthetic Image Count)".

(b) Sample heatmaps for both real and augmented data.

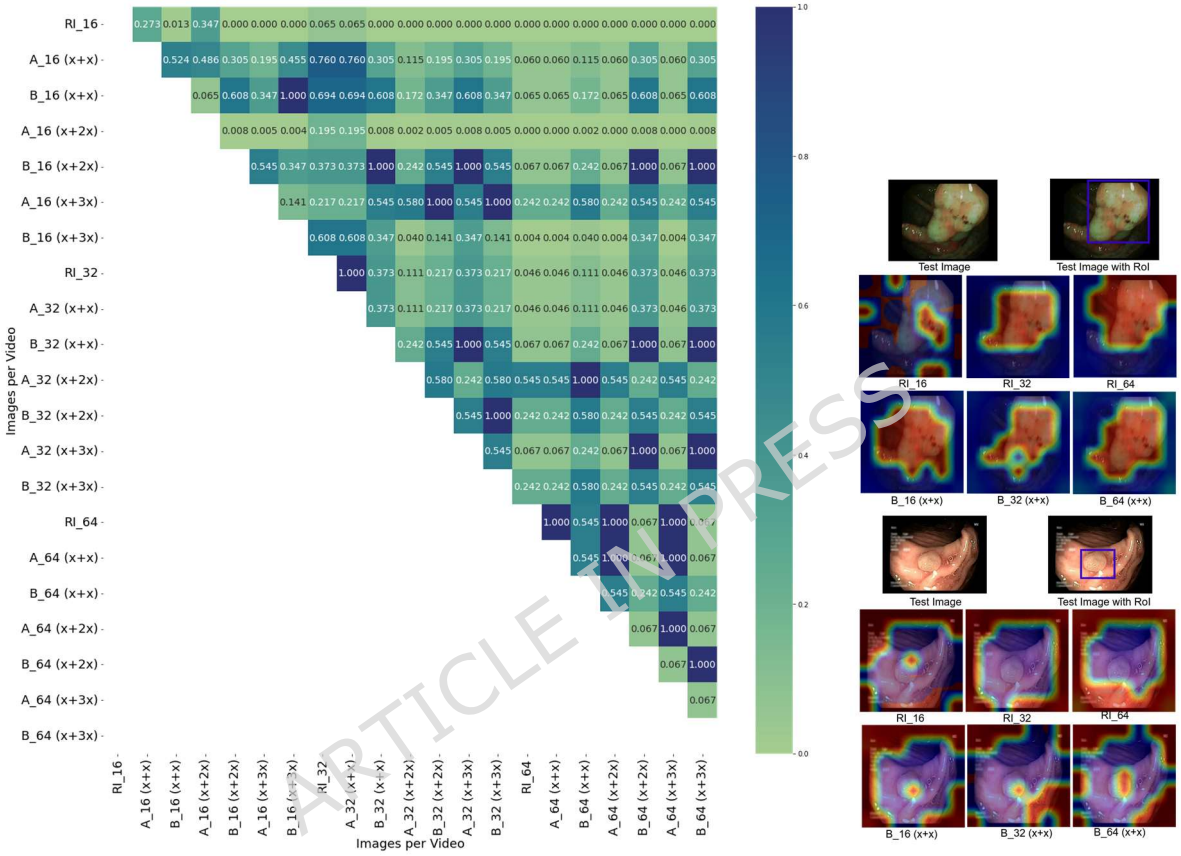
Figure 9. Video-wise outcomes obtained using WLI images and the associated p-values and heatmaps. RI stands for real images. The RIs are taken from the ISIT-UMR[9] and SUN Database[10,11].

development of this direct relationship, the model sometimes recollects "WLI" characteristics when it assigns more weightage to the tokens *"good-quality, clear"* in the given text prompt *"colonoscopy image with adenomatous polyp, narrow band imaging, good-quality, clear"* or *"colonoscopy image with hyperplastic polyp, narrow band imaging, good-quality, clear"*. Consequently, some samples present a combination of WLI and NBI images or complete WLI images.

To overcome this issue, we assigned more weight to the tokens representing *"colonoscopy image with 'y' polyp, narrow band imaging"* where 'y' can be hyperplastic/adenomatous. This modification resulted in better qualitative outcomes, as depicted in Fig. 12a to Fig. 12d. However, the related quantitative outcomes are reduced in most of the cases covered by Fig. 12e to Fig. 12g. One possible reason for such a decline in results could be due to a lack of pathology-specific features when images are generated using a weighted approach. Such outcomes signify that slight modifications in the text prompt could impact synthetic images' characteristics and visual properties.

Discussion and Limitations

In this work, our proposed text-controlled polyp generation approach presented a novel technique to generate polyps with different pathologies, which is an underexplored research area in the literature. An exhaustive investigation is conducted to study the training process of diffusion models using iteration-wise t-SNE plots, confusion matrices and quality assessment metrics. Considering the visual outcomes of t-SNE plots and the quantitative outcomes of the classification test, it can be inferred that during the training phase, our diffusion based model gradually performs well with increasing iterations; however, it starts to degrade after a certain extent. Such a trend indicates that over-training of a diffusion model could result in synthetic images with unacceptable characteristics.



(a) p-value obtained using two-tailed t-test for statistical significance analysis. The values are rounded off to 3 decimal places. The label names used in rows and columns can be read as "Text Prompt_Sample Count per Video (Real Image Count + Synthetic Image Count)".

(b) Sample heatmaps for both real and augmented data.

Figure 10. Video-wise outcomes obtained using NBI images and the associated p-values and heatmaps. RI stands for real images. The RIs are taken from the ISIT-UMR[9] and SUN Database[10,11].

Following this assessment, a standard classification test is conducted utilizing pre-trained (on real images) models, including DenseNet-201 and DenseNet-121. This aims to assess the clinical relevance of synthetic images obtained in different iterations by validating if a model trained using real images could identify the synthetic images during the inference phase. These results, along with the outcomes of KID and t-SNE plots assisted in selecting the best training iteration of our proposed diffusion model. In Stage-I, the F1-score improved with every iteration, contradicting both the qualitative results and the KID score. The t-SNE plots suggested that such a scenario occurred because, despite unsatisfactory image generation, the separation between the two classes (polyp and non-polyp) was sufficient to achieve a significantly high F1-score. However, the lack of overlap with their real counterparts resulted in unsatisfactory KID scores at higher iterations. Therefore, our study reveals that the clinical relevance of synthetic images needs to be evaluated using different standard

metrics. A similar approach was followed in Stage-II, yielding more convincing outcomes with fewer contradictions. However, a common observation in both stages was that the synthetic image data points began to deviate from the real data points after certain training iterations.

We further validated the pathological content of the synthetic images by using the generated images to augment the real images of a public dataset. This augmented dataset is then used for the downstream task of binary classification on adenomatous/hyperplastic polyps using EfficientNet-B0. We compared our augmentation approach with the baseline (only real data) using different proportions of the dataset. The best result reported an increase of 7.91% (0.6983±0.039 vs. 0.6192±0.073). With a similar comparison approach (using different data proportions), we also examined the synthetic images generated using different variations of text prompts. It is observed that the text prompts formulated using the cross-class label concept outperformed those without such labels in most of the cases (for both NBI and WLI). In addition to frame-level analysis,

we conducted video-level investigations. The associated quantitative results are supported by a statistical significance test (two-tailed t-test) and heatmaps. At video-level analysis, a statistically significant difference can be observed in favor of

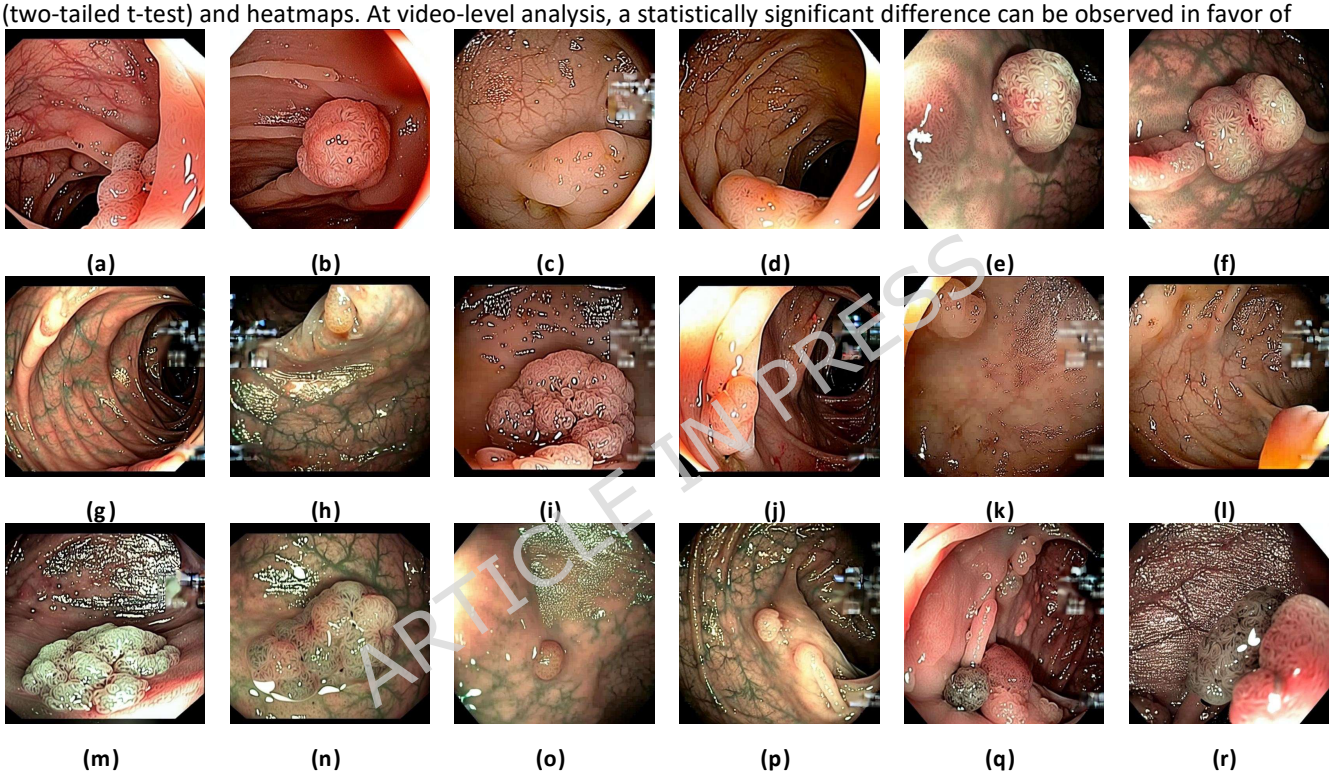


Figure 11. Sample generated images depicting (a)-(b) adenomatous polyp in WLI (using text prompt B) (c)-(d) hyperplastic polyp in WLI (using text prompt B), (e)-(f) adenomatous polyp in NBI (using text prompt B), (g)-(h) hyperplastic polyp in NBI (using text prompt B), (i)-(j) adenomatous polyp in WLI (using text prompt A), (k)-(l) hyperplastic polyp in WLI (using text prompt A), (m)-(n) adenomatous polyp in NBI (using text prompt A), (o)-(p) hyperplastic polyp in NBI (using text prompt A), and (q)-(r) shows some undesired images generated using text prompt B with NBI.

cross-class label concept. Although the text prompts without cross-class labels achieved the best outcomes for NBI cases, the overall performance of both text prompts was similar, and the difference was not statistically significant. In addition, we explored the concept of weighted text prompts and presented both qualitative outcomes and quantitative analysis through box plots. Furthermore, we examined a crucial scenario where one pathology class (adenomatous) was represented entirely by synthetic samples, while the other class (hyperplastic) contained only real images. As demonstrated in Fig. 12h and 12i, the classifier performance degrades compared to the baseline where both classes use real data. This degradation rate increases with an increase in the total synthetic sample count, indicating that simply enlarging the synthetic set introduces additional noise when not supported by real examples from that class. These results suggest that synthetic data can provide useful complementary diversity when real samples are limited, but cannot fully substitute for real data when used in

isolation. It is noteworthy that, to the best of our knowledge, this is the first work that focused on text-controlled pathology-based polyp generation and also introduced the concept of cross-class label learning.

Limitations and Future Work: It is evident from the above-mentioned analysis that our approach achieved visually appealing outcomes followed by promising quantitative results. However, our model PathoPolyp-Diff also has some limitations. In a few cases, utilizing the cross-class labels resulted in undesired qualitative outcomes due to limited generation control. When the weighted mechanism was incorporated, it improved the qualitative results but at the expense of degraded quantitative outcomes. From a clinical perspective, such undesired outcomes could be critical, as pathologically unrealistic images, if used for training diagnostic models, can lead to inaccurate predictions and misdiagnosis. Further exploration is needed in such cases. For instance, an auxiliary classifier trained separately on good quality and clean samples can be used to guide the diffusion process. In addition, some parameters like guidance and temperature scale can be adjusted for improvement. We also highlight that while our study provides empirical analysis of prompt sensitivity, including negative prompts and weighted mechanism, an exhaustive analysis of prompt sensitivity across all possible linguistic variations represents an important direction for future work. During iteration-wise analysis, we observed that synthetic features begin to deviate from the real data manifold after 6,000-10,000 iterations, resulting in distorted textures and unusual color patterns. Addressing this deviation remains a key priority for future work. Specifically, we intend to implement an auxiliary feedback loop utilizing a pre-trained diagnostic classifier. By penalizing outputs that lack clinical significance during the denoising process, we aim to extend the model's

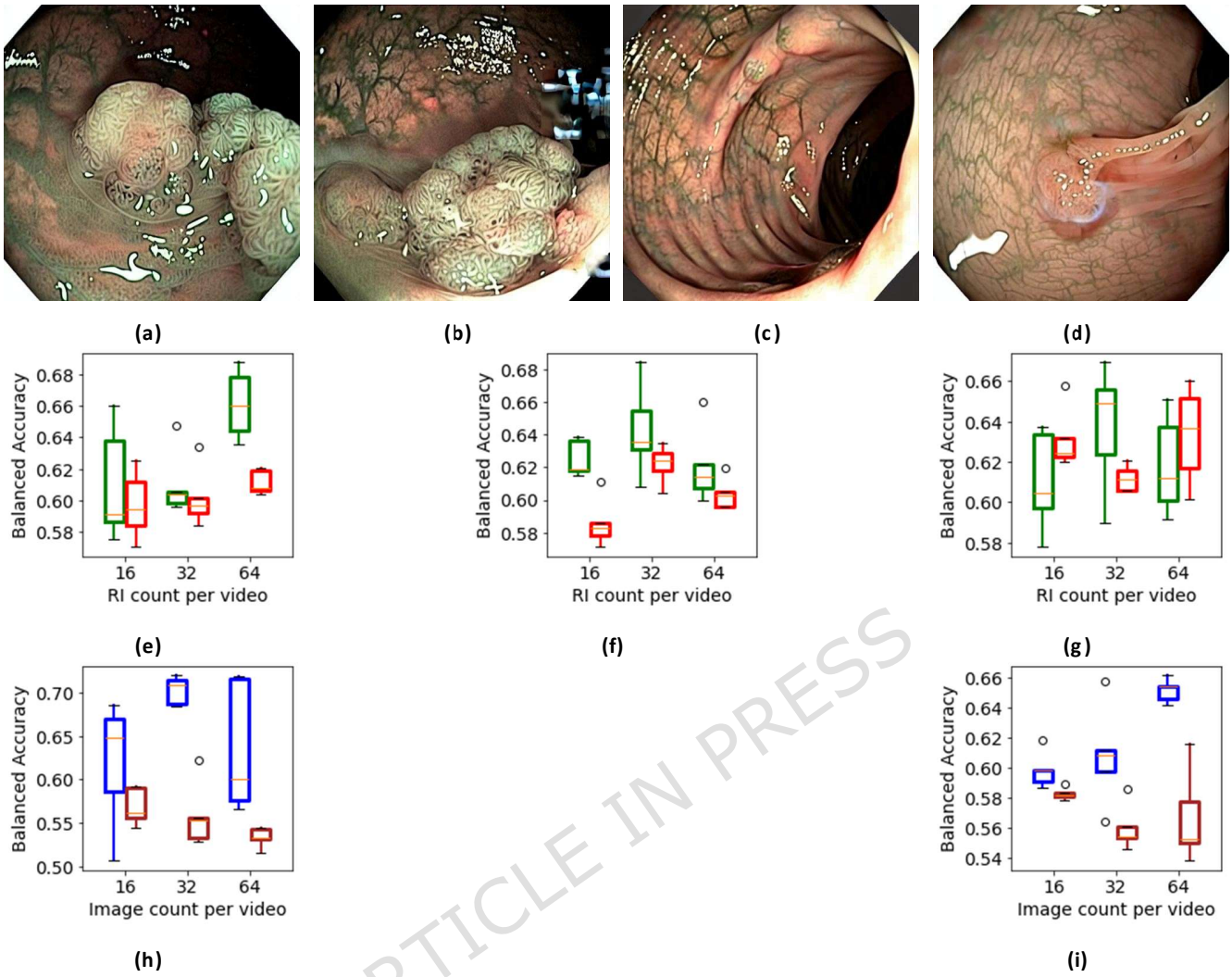


Figure 12. Sample generated images depicting (a)-(b) adenomatous polyps in NBI, and (c)-(d) hyperplastic polyps in NBI using weighted control mechanism. Boxplots (e) to (g) demonstrate the comparison between text prompt B with and without a weighted control mechanism

when synthetic images are added in equal proportion or twice or thrice in proportion to the real images, respectively. The former and latter text prompts are denoted by green and red color, respectively. RI stands for Real Images. Boxplots (h) and (i) compare the baseline condition (all real images) with a mixed condition in which one class is fully synthetic and the other is real (adenoma synthetic, hyperplastic real). The baseline is shown in blue, and the mixed condition is shown in brown.

stability and further enhance the informative gain of the synthetic data.

Regarding computational considerations, our method, being based on diffusion models, inherently involves a complex architecture and computationally intensive processes. Given the current design and objectives of the model, it may not be well-suited for deployment on low-end devices at this time. In a clinical setting, this issue can lead to local adaptation challenges where fine-tuning on a local dataset could be desirable for a specific task. The other consequences can include slow inference speed during emergency scenarios and high memory demands in remote healthcare facilities. To address this issue, some techniques, such as model quantization, pruning, and knowledge distillation, can be explored. Despite the above limitations, we successfully developed a novel diffusion-based technique to generate a diverse set of polyps using cross-class labels; we also provide a roadmap for the research community to build upon our work, extending the synthetic polyp dataset and experimenting with text prompts to enhance overall outcomes. Additionally, as already discussed, other possible exploration directions could include model simplifications, quantization, or developing lightweight variants that

maintain performance while being more resource-efficient. Further, clinical validation by experts would help to rigorously assess the quality and realism of the generated images, validating their acceptability in a clinical setting.

Conclusion

In this paper, we developed a novel diffusion based model, PathoPolyp-Diff, which generates a visually appealing diverse set of polyp images in a two-stage process. This set of polyps covers multiple categories, including pathology (adenomatous/hyperplastic), imaging modality (NBI/WLI), and quality (informative/uninformative). In addition, we introduced the concept of cross-class label learning, which allows the generation of diversified polyps exhibiting features from other classes

without requiring additional labels. This text-controlled image generation is validated by performing a downstream binary classification task on adenomatous and hyperplastic polyp classes with NBI and WLI modalities. Further, we studied the impact of text prompt variations on colonoscopy image generation by incorporating a weighted control mechanism. The results are analyzed at both frame and video level outcomes with a statistical significant test. Our research also reveals the significance of text prompt formulation on synthetic images' visual and clinical properties. The proposed approach achieved an improvement of 7.91% and 18.33% in balanced accuracy when used for augmentation and cross-class label learning, respectively.

Data Availability

All data used in this study are publicly available or accessible on request. The SUN Database details are available at http://sundatabase.org/ (Ref. 26, 27), and it requires a prior request from the respective authors for access, while the ISIT-UMR Colonoscopy Dataset is directly available at http://www.depeca.uah.es/colonoscopy_dataset/ (Ref.28). The quality-based annotations can be obtained on request (Ref. 29). This study did not involve humans or any living organism subjects.

References

1. Organization, W. H. Colorectal cancer fact sheets. https://www.who.int/news-room/fact-sheets/detail/colorectal-cancer (2023).
2. Cheng, W. *et al.* Narrow-band imaging endoscopy is advantageous over conventional white light endoscopy for the diagnosis and treatment of children with peutz-jeghers syndrome. *Medicine* 96 (2017).
3. Zhu, J., Yang, G. & Lio, P. How can we make gan perform better in single medical image super-resolution? a lesion focused multi-scale approach. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, 1669–1673 (2019).
4. Bhadra, S., Zhou, W. & Anastasio, M. A. Medical image reconstruction with image-adaptive priors learned by use of generative adversarial networks. In *Medical Imaging 2020: Physics of Medical Imaging*, vol. 11312, 206–213 (2020).
5. Xun, S. *et al.* Generative adversarial networks in medical image segmentation: A review. *Comput. biology medicine* 140, 105063 (2022).
6. Sharma, V., Bhuyan, M. & Das, P. K. Can adversarial networks make uninformative colonoscopy video frames clinically informative?(student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, 16322–16323 (2023).
7. Wolleb, J., Bieder, F., Sandkühler, R. & Cattin, P. C. Diffusion models for medical anomaly detection. In *International Conference on Medical image computing and computer-assisted intervention*, 35–45 (2022).
8. Rahman, A., Valanarasu, J. M. J., Hacihaliloglu, I. & Patel, V. M. Ambiguous medical image segmentation using diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11536–11546 (2023).
9. Mesejo, P. *et al.* Computer-aided classification of gastrointestinal lesions in regular colonoscopy. *IEEE transactions on medical imaging* 35, 2051–2063 (2016).
10. Itoh, H. *et al.* Sun colonoscopy video database. http://amed8k.sundatabase.org/ (2020).
11. Misawa, M. *et al.* Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video). *Gastrointest. endoscopy* 93, 960–967 (2021).

12. Machácek,˘ R. *et al.* Mask-conditioned latent diffusion for generating gastrointestinal polyp images. *arXiv preprint arXiv:2304.05233* (2023).

13. Pishva, A. K., Thambawita, V., Torresen, J. & Hicks, S. A. Repolyp: A framework for generating realistic colon polyps with corresponding segmentation masks using diffusion models. In *2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)*, 47–52 (2023).

14. Shin, Y., Qadir, H. A. & Balasingham, I. Abnormal colon polyp image synthesis using conditional adversarial networks for improved detection performance. *IEEE Access* 6, 56007–56017 (2018).

15. Qadir, H. A., Balasingham, I. & Shin, Y. Simple u-net based synthetic polyp image generation: Polyp to negative and negative to polyp. *Biomed. Signal Process. Control.* 74, 103491 (2022).

16. Fagereng, J. A. *et al.* Polypconnect: Image inpainting for generating realistic gastrointestinal tract images with polyps. In *2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS)*, 66–71 (2022).

17. Sasmal, P., Bhuyan, M., Sonowal, S., Iwahori, Y. & Kasugai, K. Improved endoscopic polyp classification using gan generated synthetic data augmentation. In *2020 IEEE Applied Signal Processing Conference (ASPCON)*, 247–251 (2020).

18. Adjei, P. E., Lonseko, Z. M., Du, W., Zhang, H. & Rao, N. Examining the effect of synthetic data augmentation in polyp detection and segmentation. *Int. J. Comput. Assist. Radiol. Surg.* 17, 1289–1302 (2022).

19. He, F. *et al.* Colonoscopic image synthesis for polyp detector enhancement via gan and adversarial training. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 1887–1891 (2021).

20. Sams, A. & Shomee, H. H. Gan-based realistic gastrointestinal polyp image synthesis. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, 1–4 (2022).

21. Golhar, M., Bobrow, T. L., Ngamruengphong, S. & Durr, N. J. Gan inversion for data augmentation to improve colonoscopy lesion classification. *arXiv preprint arXiv:2205.02840* (2022).

22. Bhamre, N. V., Sharma, V., Iwahori, Y., Bhuyan, M. & Kasugai, K. Colonoscopy polyp classification adding generated narrow band imaging. In *International Conference on Computer Vision and Image Processing*, 322–334 (2022).

23. Du, Y. *et al.* Arsdm: Colonoscopy images synthesis with adaptive refinement semantic diffusion models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 339–349 (2023).

24. Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695 (2022).

25. Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. *Adv. neural information processing systems* 33, 6840–6851 (2020).

26. Kingma, D. P., Welling, M. *et al.* An introduction to variational autoencoders. *Foundations Trends Mach. Learn.* 12, 307–392 (2019).

27. Ronneberger, O., Fischer, P. & Brox, T. 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*, 234–241 (2015).

28. Radford, A. *et al.* Learning transferable visual models from natural language supervision. In *Proceedings of the International conference on machine learning*, 8748–8763 (2021).

29. Sharma, V. *et al.* A multi-scale attention framework for automated polyp localization and keyframe extraction from colonoscopy videos. *IEEE Transactions on Autom. Sci. Eng.* (2023).

30. Binkowski,´ M., Sutherland, D. J., Arbel, M. & Gretton, A. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401* (2018).

31. Jha, D. *et al.* Gastrovision: A multi-class endoscopy image dataset for computer aided gastrointestinal disease detection. In *Workshop on Machine Learning for Multimodal Healthcare Data*, 125–140 (Springer, 2023).

32. Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. & Smola, A. A kernel two-sample test. *The journal machine learning research* 13, 723–773 (2012).

33. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708 (2017).

34. Zhang, R., Isola, P., Efros, A. A., Shechtman, E. & Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595 (2018).

35. Tan, M. & Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the International conference on machine learning*, 6105–6114 (2019).

36. Sharma, V. *et al.* Controlpolypnet: Towards controlled colon polyp synthesis for improved polyp segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2325–2334 (2024).

37. Sagers, L. W. *et al.* Augmenting medical image classifiers with synthetic data from latent diffusion models. *arXiv preprint arXiv:2308.12453* (2023).

Acknowledgements and Funding Information

V.S. is supported by the INSPIRE fellowship (IF190362), DST, Govt. of India. D.J. is supported by the NIH funding: R01-CA246704, R01-CA240639, U01 DK127384-02S1, and U01-CA268808. V.S. would like to thank Dr. Abhishek for reviewing the article.

Author contributions statement

V.S. and D.J. conceived the study. M.K.B, P.K.D and U.B designed the article. V.S. conducted experiments, designed illustrations and substantially contributed to the writing of the article. D.J. analyzed the results and participated in manuscript

writing. All authors reviewed the manuscript and provided critical feedback.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to V.S.