

# CCIS-DIFF: A GENERATIVE MODEL WITH STABLE DIFFUSION PRIOR FOR CONTROLLED COLONOSCOPY IMAGE SYNTHESIS

Yifan Xie<sup>1,4</sup>, Jingge Wang<sup>2,3</sup>, Tao Feng<sup>2</sup>, Fei Ma<sup>4\*</sup>, Yang Li<sup>2,3\*</sup>

<sup>1</sup>School of Software Engineering, Xi’an Jiaotong University, Xi’an, China

<sup>2</sup>Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China

<sup>3</sup>Shenzhen Key Laboratory of Ubiquitous Data Enabling, Shenzhen, China

<sup>4</sup>Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen, China

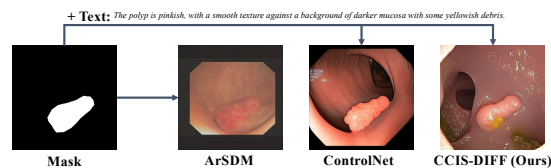
## ABSTRACT

Colonoscopy is crucial for identifying adenomatous polyps and preventing colorectal cancer. However, developing robust models for polyp detection is challenging by the limited size and accessibility of existing colonoscopy datasets. While previous efforts have attempted to synthesize colonoscopy images, current methods suffer from instability and insufficient data diversity. Moreover, these approaches lack precise control over the generation process, resulting in images that fail to meet clinical quality standards. To address these challenges, we propose **CCIS-DIFF**, a Controlled generative model for high-quality Colonoscopy Image Synthesis based on a **Diffusion** architecture. Our method offers precise control over both the spatial attributes (polyp location and shape) and clinical characteristics of polyps that align with clinical descriptions. Specifically, we introduce a blur mask weighting strategy to seamlessly blend synthesized polyps with the colonic mucosa, and a text-aware attention mechanism to guide the generated images to reflect clinical characteristics. Notably, to achieve this, we construct a new multi-modal colonoscopy dataset that integrates images, mask annotations, and corresponding clinical text descriptions. Experimental results demonstrate that our method generates high-quality, diverse colonoscopy images with fine control over both spatial constraints and clinical consistency, offering valuable support for downstream segmentation and diagnostic tasks.

**Index Terms**— Colonoscopy Image Synthesis, Controlled Synthesis, Stable Diffusion

## 1. INTRODUCTION

Colonoscopy is an essential tool for detecting adenomatous polyps and reducing rectal cancer mortality rates [1]. However, training models for automatic polyp detection is challenging due to the small scale of available colonoscopy



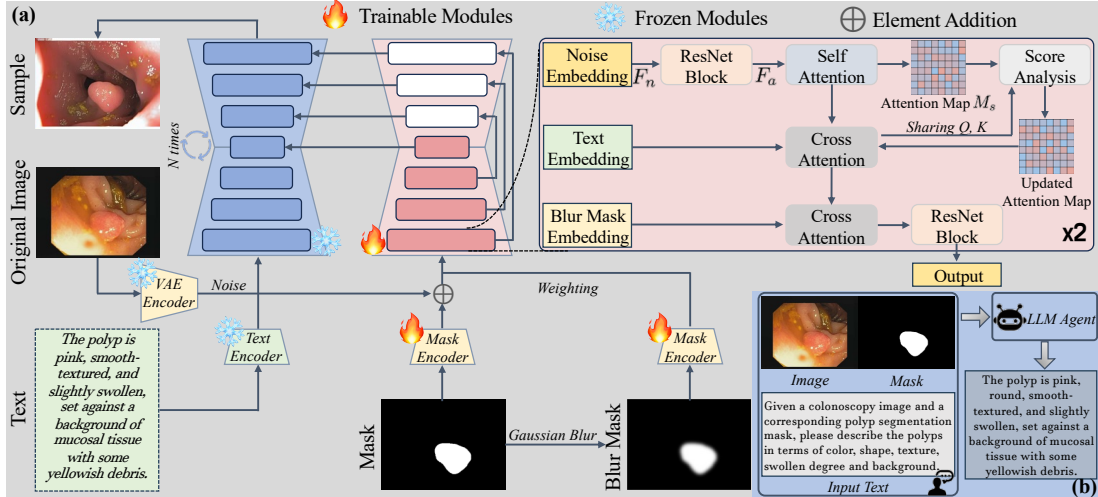
**Fig. 1.** In contrast to ArSDM [2] and ControlNet [3], CCIS-DIFF can utilize not only mask but also textual description to generate high-fidelity, text-consistent colonoscopy images.

datasets, making it difficult to have sufficient robustness and generalization that meet real-world clinical demands.

To address this problem, previous methods [4, 5, 2, 6] primarily relied on generative adversarial networks or diffusion models to synthesize more colonoscopy images. Although these efforts aim to address the data scarcity problem, they struggle to generate a sufficiently diverse and high-quality image, and the generation process lacks adequate control, leading to images that fail to meet clinical requirements for practical use. As illustrated in Fig. 1, ArSDM [2] only utilizes the mask to synthesize the colonoscopy image and the generated image is of poor quality and contains noise. Meanwhile, large-scale text-to-image (T2I) diffusion models such as Stable Diffusion [7] and DALL-E [8] have demonstrated remarkable capabilities in generating images from various prompts. This raises an important question: Can colonoscopy images be generated in a controlled manner using a pre-trained large-scale T2I model? In response, we present an innovative generative method to synthesize high-quality colonoscopy images in a controlled manner. The main contributions of this paper can be summarized as follows:

- We propose a novel generative model, named CCIS-DIFF, which offers fine control over both the spatial attributes and clinical characteristics of polyps, enabling more clinically consistent image synthesis for practical use.
- We introduce the blur mask weighting strategy to ensure seamless integration of synthesized polyps with the colonic mucosa, along with a text-aware attention mechanism that incorporates textual information into the generation process, to enable customization of polyp images.

\*Corresponding email: mafei@gml.ac.cn, yangli@sz.tsinghua.edu.cn



**Fig. 2.** (a) The overall architecture of CCIS-DIFF. The parameters of the frozen modules refer to Stable Diffusion v1.5 [7]. The trainable modules incorporate a text-aware attention mechanism to address the issue of neglecting text prompts. Additionally, we introduce a blur mask weighting strategy to ensure seamless integration of synthesized polyps with the colonic mucosa. (b) The construction pipeline of the multi-modal colonoscopy dataset.

- We design the first multi-modal colonoscopy dataset that uniquely combines colonoscopy images, segmentation masks, and clinical text descriptions, providing precise alignment between visual data and clinical information for improved training and evaluation.

## 2. METHOD

### 2.1. Multi-Modal Colonoscopy Dataset

A key challenge for current diffusion models in colonoscopy image synthesis [2, 6] is the absence of a dataset with consistent spatial characteristics of polyps and their corresponding clinical textual descriptions. Existing datasets, such as EndoScene [9], CVC-ClinicDB [10], and Kvasir [11], primarily contain image-mask pairs that focus on polyp regions. However, these datasets do not include detailed clinical text descriptions, limiting their usefulness for training models that aim to generate clinically accurate and diverse synthetic images. To address this limitation, we built a new multi-modal colonoscopy dataset based on existing datasets, containing three key components: colonoscopy images, segmentation masks, and clinical text descriptions. Such a dataset is vital for fine-tuning pretrained diffusion models, allowing them to adapt effectively to controlled colonoscopy image synthesis.

An overview of our dataset construction process is shown in Fig. 2 (b). To generate accurate captions that accurately reflect both spatial constraints and relevant clinical characteristics, we construct the open-source LLaMA [12] large language model (LLM) agent. This agent takes a textual prompt and an image-mask colonoscopy pair to create captions for both the foreground and background. To increase textual diversity, the LLM incorporates different aspects like

color, shape, texture, and swelling. As a result, we develop a multi-modal colonoscopy dataset consisting of triplets of colonoscopy images, mask images, and their corresponding textual descriptions which provides the necessary foundation for generating images with improved realism and variability.

### 2.2. CCIS-DIFF Architecture

#### 2.2.1. Overview

The overview architecture of our CCIS-DIFF is presented in Fig. 2 (a). Using our multi-modal colonoscopy dataset, we provide the original colonoscopy image  $I$ , the mask image  $M$ , and the corresponding text description  $T$ . Each of these components passes through its respective encoder, noting that the text encoder is frozen. Additionally, we develop a blur mask to ensure that the generated polyp integrates seamlessly with the background, the detailed description will be illustrated in Sec. 2.2.2.

Based on ControlNet [3], we adopt the trainable diffusion branch and implement the zero convolution strategy to protect this branch by eliminating random noise as gradients in the initial training steps. This structure, when applied to large models like Stable Diffusion [7], enables the frozen parameters to preserve the integrity of the production-ready model that has been trained on billions of images. Meanwhile, the trainable diffusion branch leverages this large-scale pretrained model to establish a robust backbone for managing multi-modal input conditions. Furthermore, to address the issue of neglecting text prompt and effectively incorporating textual information into the generation process, we incorporate a text-aware attention mechanism (Sec. 2.2.3) in the trainable diffusion branch.

### 2.2.2. Blur Mask Weighting Strategy

The purpose of the blur mask weighting strategy is to ensure that the generated polyp is seamlessly integrated with the background. To achieve this, we apply a Gaussian blur operation  $\sigma$  to the mask image  $M$ , softening the transition between the polyp mask region and the background. We utilize two separate mask encoders to extract features from the mask and the blurred mask, each with non-shared parameters. Subsequently, a weighting matrix  $M_w$  is introduced to balance the two mask branches, and  $M_w$  is learned using a three-layer MLP. Thus the blur mask embedding  $F_b$  can be constructed as:

$$F_b = M_w \odot E_m(\sigma(M)), \quad (1)$$

where  $\odot$  is the Hadamard product and  $E_m$  denotes the mask encoder.

### 2.2.3. Text-Aware Attention Mechanism

In our experiments, we observed that existing methods, such as ControlNet [3], often overlook the text prompt and rely more heavily on the mask image. We hypothesize that this visual dominance over the text prompt arises from the text-free nature of the self-attention layers. To address this issue, we propose a text-aware attention mechanism that leverages the cross-attention matrix to regulate the output of the self-attention.

Given the noise embedding  $F_n$ , it is first passed through a ResNet block, as referenced in ControlNet [3]. Following this, we obtain the attention input tensor  $F_a \in \mathbb{R}^{(h \times w) \times d}$ , which is processed through projection layers to derive the queries  $Q_s$ , keys  $K_s$ , and values  $V_s$ , as well as the attention map  $M_s = \frac{Q_s K_s^T}{\sqrt{d}} \in \mathbb{R}^{hw \times hw}$ . To reduce the strong influence of the mask image, we adjust the attention scores of the prompt text. Specifically, we begin by constructing the cross-attention similarity matrix:

$$M_c = \text{SoftMax}(Q_c K_c^T / \sqrt{d}), \quad (2)$$

where  $Q_c \in \mathbb{R}^{(h \times w) \times d}$ ,  $K_c \in \mathbb{R}^{l \times d}$  represent the query and key tensors from their corresponding cross-attention layers, and  $l$  denotes the number of tokens in the text prompt. For each pixel  $j$ , we define its similarity to the text prompt by summing its similarity scores with the embedding indexed by  $M_c$ . We then apply a clamp operation to normalize the scores  $s_j$  within the range  $[0, 1]$ :

$$s_j = \text{Norm}(\text{Sum}_j(M_c)). \quad (3)$$

By calculating the scores for each pixel, we can obtain the final text-aware map  $S$ . Subsequently, we utilize  $S$  to compute the updated attention map and refine the self-attention process.

Finally, the output of the cross-attention mechanism interacts once more with the blur mask embedding, generating the final output embedding through a ResNet block.

**Table 1.** The quantitative results of the colonoscopy image synthesis. The boldface indicates the best performance.

Method	FID ↓	CLIP-score ↑	CLIP-image ↑
ControlNet	87.32	30.78	87.09
Uni-ControlNet	92.74	30.57	86.46
Ours	<b>71.73</b>	<b>31.96</b>	<b>88.70</b>

### 2.2.4. Training and Inference

During the training phase, the diffusion models start with an original image  $I$  and progressively add noise to create a noisy embedding  $I_t$ , where  $t$  indicates the number of times noise is added. The models learn a network  $\epsilon_\theta$  to predict the noise added to the noisy image  $I_t$ , based on a set of conditions that include the time step  $t$ , a text prompt  $T$ , and a mask image  $M$ . The overall loss function  $\mathcal{L}$  for the entire diffusion model is represented as:

$$\mathcal{L} = \mathbb{E}_{I,t,T,M,\epsilon \sim \mathcal{N}(0,1)} [\|\epsilon - \epsilon_\theta(I_t, t, T, M)\|_2^2]. \quad (4)$$

Additionally, we randomly replace half of the text prompts with empty strings. This strategy enhances the model’s ability to directly recognize the semantics of the input mask image, serving as a substitute for the text prompt.

During the inference phase, the diffusion models automatically generate noise and produce the final colonoscopy images based on a text prompt and a mask image. Simultaneously, we utilize Classifier-Free Guidance (CFG) [13] to enhance the sampling process.

## 3. EXPERIMENT AND RESULTS

### 3.1. Implementation Details

All methods were implemented in PyTorch on a single NVIDIA A100 40G GPU. We utilized the pre-trained Stable Diffusion v1.5 [7] model, following ControlNet [3], to replicate its UNet encoder as the trainable copy encoder. The batch size was configured to 4, and the learning rate was set to 1e-5. For inference, we used a default CFG scale of 7.0. The DDIM sampler was employed, using 20 steps to sample each image.

### 3.2. Baseline Algorithms and Evaluation Metrics

For the colonoscopy image synthesis task, we compared our method with ControlNet [3] and Uni-ControlNet [14], all of which were trained on our multi-modal colonoscopy dataset. Following [3], we assessed performance using FID, CLIP-score, and CLIP-image (measuring the similarity between the generated image and the reference image).

For the downstream polyp segmentation task, we utilized PraNet [1] and Polyp-PVT [15] as baseline segmentation

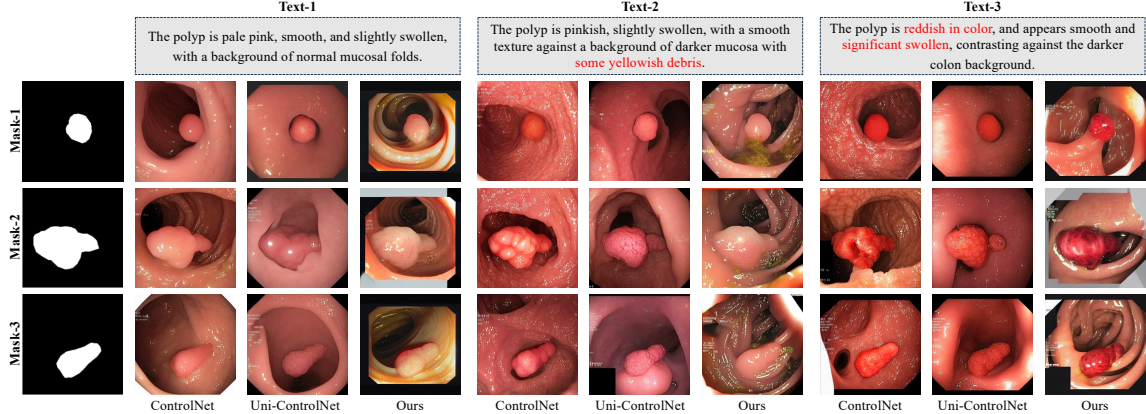


Fig. 3. Qualitative comparison of colonoscopy image synthesis.

Table 2. User study. The rating scale ranges from 1 to 5, with higher numbers indicating better performance.

Method	Image Fidelity	Mask Accuracy	Text Accuracy
ControlNet	3.852	3.962	3.918
Uni-ControlNet	2.904	2.966	3.392
Ours	<b>4.266</b>	<b>4.466</b>	<b>4.518</b>

Table 3. Comparisons of different settings applied on two polyp segmentation baselines across three public datasets.

Settings	EndoScene		CVC-ClinicDB		Kvasir	
	mDice	mIoU	mDice	mIoU	mDice	mIoU
PraNet	86.1	79.1	90.8	86.1	89.2	84.0
PraNet + ControlNet	85.4	77.7	90.1	85.1	90.6	85.2
PraNet + Uni-ControlNet	85.9	78.2	89.5	85.1	87.9	82.5
PraNet + Ours	<b>88.5</b>	<b>81.0</b>	<b>92.8</b>	<b>88.0</b>	<b>92.8</b>	<b>87.1</b>
PVT	88.2	81.2	92.8	87.6	91.5	86.8
PVT + ControlNet	88.3	81.4	92.7	88.2	90.3	84.8
PVT + Uni-ControlNet	85.8	79.6	89.3	83.8	86.5	80.7
PVT + Ours	<b>89.3</b>	<b>82.1</b>	<b>93.5</b>	<b>88.4</b>	<b>92.4</b>	<b>87.0</b>

models with their default settings, and assessed their performance using mean Dice (mDice) and mean Intersection over Union (mIoU) metrics.

### 3.3. Results

The quantitative results of the colonoscopy image synthesis are illustrated in Table 1. Our method outperforms others in both visual quality and text alignment. Furthermore, we present a comparison of our method with others in Fig. 3. The results show that the polyp region generated by our method is more consistent with the mask image, and the overall information aligns more closely with the text prompt.

To conduct a more comprehensive evaluation, we implemented a user study questionnaire. We selected 30 groups of images for comparison and invited 15 clinical experts from the First Affiliated Hospital of Sun Yat-sen University to participate in the survey. Participants were required to rate the synthesized images based on three criteria: image fidelity, mask accuracy, and text accuracy. The average scores for each criterion are presented in Table 2. Our method outperforms

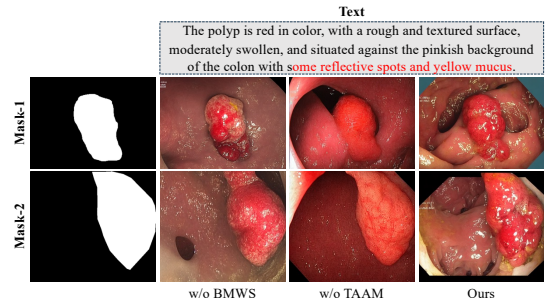


Fig. 4. Visualization of the ablation experiments.

the other methods in all aspects.

To further verify the effectiveness of our synthetic images, we evaluated our method against state-of-the-art methods for the polyp segmentation task. We generated the same number of samples as the diffusion training set using the original masks, which we then combined to create a new downstream training dataset. The experimental results presented in Table 3 highlight the effectiveness of our method in training improved downstream models that achieve superior performance.

We also conduct ablation visualization experiments to demonstrate the effectiveness of the Blur Mask Weighting Strategy (BMWS) and the Text-Aware Attention Mechanism (TAAM). The results are illustrated in Fig. 4. The combination of BMWS and TAAM significantly enhances the quality of the synthesized images and text alignment.

## 4. CONCLUSION

In this paper, we present CCIS-DIFF, a generative model that leverages a Stable Diffusion prior for the controlled colonoscopy image synthesis. We begin by developing a blur mask weighting strategy to seamlessly integrate the generated polyp with the colonic mucosa, along with a text-aware attention mechanism to address the issue of neglecting text prompt. Additionally, we introduce a multi-modal colonoscopy dataset created by large language models. Extensive experiments across various settings demonstrate the superior performance of CCIS-DIFF.

## 5. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access. Ethical approval was not required as confirmed by the license attached with the open access data.

## 6. REFERENCES

- [1] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao, “Pranet: Parallel reverse attention network for polyp segmentation,” in *International conference on medical image computing and computer-assisted intervention*. Springer, 2020, pp. 263–273.
- [2] Yuhao Du, Yuncheng Jiang, Shuangyi Tan, Xusheng Wu, Qi Dou, Zhen Li, Guanbin Li, and Xiang Wan, “Ar-sdm: colonoscopy images synthesis with adaptive refinement semantic diffusion models,” in *International conference on medical image computing and computer-assisted intervention*. Springer, 2023, pp. 339–349.
- [3] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala, “Adding conditional control to text-to-image diffusion models,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023*, pp. 3836–3847.
- [4] Yuhui Ma, Yonghuai Liu, Jun Cheng, Yalin Zheng, Morteza Ghahremani, Honghan Chen, Jiang Liu, and Yitian Zhao, “Cycle structure and illumination constrained gan for medical image enhancement,” in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23*. Springer, 2020, pp. 667–677.
- [5] Jiabo Xu, Saeed Anwar, Nick Barnes, Florian Grimpén, Olivier Salvado, Stuart Anderson, and Mohammad Ali Armin, “Ofgan: Realistic rendition of synthetic colonoscopy videos,” in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*. Springer, 2020, pp. 732–741.
- [6] Minjae Jeong, Hyuna Cho, Sungyoon Jung, and Won Hwa Kim, “Uncertainty-aware diffusion-based adversarial attack for realistic colonoscopy image synthesis,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 647–658.
- [7] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, “High-resolution image synthesis with latent diffusion models,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022*, pp. 10684–10695.
- [8] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen, “Hierarchical text-conditional image generation with clip latents,” *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, pp. 3, 2022.
- [9] David Vázquez, Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Antonio M López, Adriana Romero, Michal Drozdal, and Aaron Courville, “A benchmark for endoluminal scene segmentation of colonoscopy images,” *Journal of healthcare engineering*, vol. 2017, no. 1, pp. 4037190, 2017.
- [10] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño, “Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians,” *Computerized medical imaging and graphics*, vol. 43, pp. 99–111, 2015.
- [11] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas De Lange, Dag Johansen, and Håvard D Johansen, “Kvasir-seg: A segmented polyp dataset,” in *MultiMedia modeling: 26th international conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, proceedings, part II 26*. Springer, 2020, pp. 451–462.
- [12] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al., “Llama: Open and efficient foundation language models,” *arXiv preprint arXiv:2302.13971*, 2023.
- [13] Jonathan Ho and Tim Salimans, “Classifier-free diffusion guidance,” *arXiv preprint arXiv:2207.12598*, 2022.
- [14] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong, “Uni-controlnet: All-in-one control to text-to-image diffusion models,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [15] Bo Dong, Wenhai Wang, Deng-Ping Fan, Jinpeng Li, Huazhu Fu, and Ling Shao, “Polyp-pvt: Polyp segmentation with pyramid vision transformers,” *arXiv preprint arXiv:2108.06932*, 2021.