Structure-Consistent Joint Image-Mask Synthesis for Data-Scarce Medical Image Segmentation

Ningyan Zhang[1,*] Weiyi Zhang[2,*] Mostofa Rafid Uddin[1] Xingjian Li[1,†] Min Xu[1,†]

[1]Carnegie Mellon University [2]University of California, Berkeley

[*]Equal contribution. [†]Corresponding authors.

Abstract

Synthetic data generation is a promising strategy for medical image segmentation under limited annotation, but its usefulness depends on whether the generated image and mask remain structurally consistent as a paired training sample. Existing diffusion-based methods have shown strong potential for medical image synthesis, yet many primarily focus on mask-conditioned image generation rather than explicit image–mask pair modeling. In this work, we propose a structure-consistent framework for joint image–mask synthesis in data-scarce medical segmentation. Our method jointly models the image and mask within a single diffusion process and introduces a lightweight Feature-wise Linear Modulation (FiLM)-based mask-conditioned control branch for additional structural guidance. To enable controllable sampling, we further train a mask-only diffusion model and use its sampled pseudo masks as conditioning inputs during inference. Experiments on representative RGB and grayscale datasets show that our method performs best in the more challenging RGB setting and remains competitive in the near-ceiling grayscale setting.

1. Introduction

Medical image segmentation is a fundamental component of computer-aided diagnosis and clinical decision support, yet its progress is often constrained by limited annotated data and the high cost of acquiring pixel-level masks from domain experts [16]. Synthetic data generation has therefore attracted growing interest as a practical strategy for expanding supervision under low-data regimes [6, 7, 10–12, 17]. In parallel, diffusion models have emerged as a powerful generative paradigm in medical imaging, showing strong potential for realistic and diverse sample synthesis across a broad range of applications [3].

However, for segmentation augmentation, visual realism alone is not sufficient. Synthetic samples are only useful when the generated image and mask remain reliably aligned and can serve as trustworthy paired supervision for downstream training. Prior work has shown promising progress in controllable generation and synthetic-data augmentation for medical imaging [4, 5, 17]. Nevertheless, under limited real annotations, synthetic data must provide structurally reliable image–mask pairs with strong spatial correspondence to support downstream segmentation. This motivates a formulation that directly focuses on structure-consistent paired generation.

In this work, we formulate synthetic data generation as joint image–mask synthesis for data-scarce medical image segmentation. As illustrated in Figure 1, our framework jointly generates an image and its corresponding mask, while introducing an additional Feature-wise Linear Modulation (FiLM)-based mask-conditioned control branch to strengthen structural consistency during generation [9]. During inference, conditioning masks are provided as pseudo masks generated by a mask-only diffusion model trained on the limited real masks. We evaluate the proposed method on two public medical benchmarks, HyperKvasir and JSRT. HyperKvasir contains gastrointestinal endoscopy images with polyp annotations, while JSRT consists of chest X-ray scans for lung structure segmentation. All experiments are validated via downstream UNet++ [18] segmentation performance. Results show that our method is effective in the more challenging RGB setting and remains competitive in the near-ceiling grayscale setting.

Contributions. (1) We formulate synthetic data generation for data-scarce medical segmentation as joint image–mask synthesis, with an additional mask-conditioned control branch to strengthen structural consistency. (2) We explicitly distinguish the jointly synthesized target mask from the FiLM-based conditioning mask, which is generated by a mask-only diffusion model during inference for controllable synthesis. (3) We validate the synthesized pairs through low-data UNet++ segmentation on RGB and grayscale datasets, showing the strongest gains in the more challenging RGB setting while remaining competitive in the near-ceiling grayscale setting.

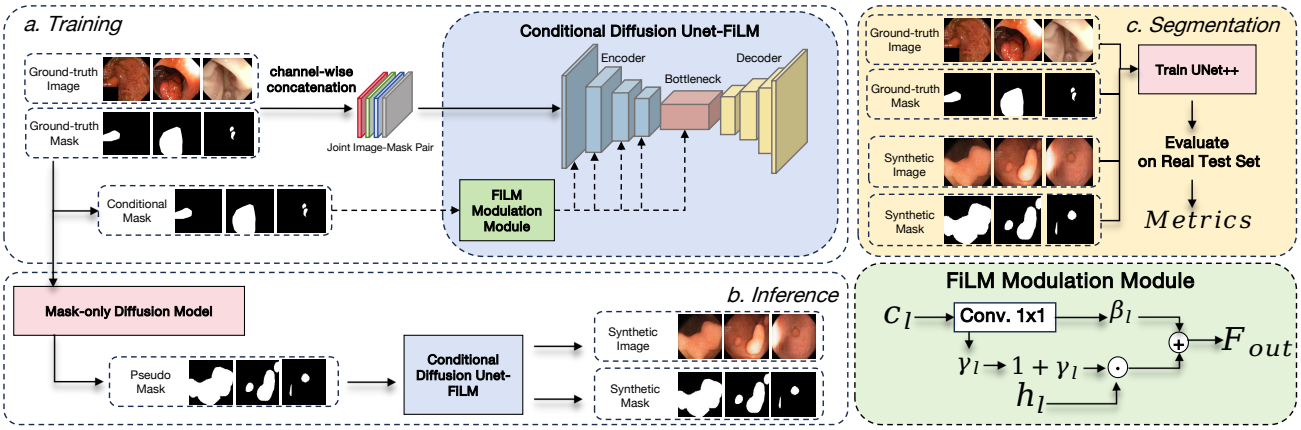

Figure 1. Overview of the proposed structure-consistent joint image–mask synthesis framework. (a) Training: the input image and target mask are concatenated as a joint diffusion target, while a conditioning mask from the same ground-truth source is separately injected through FiLM to provide structural guidance. The FiLM modulation module maps the resized conditioning mask c_l through a 1×1 convolution to generate spatial modulation parameters γ_l and β_l, which modulate the intermediate feature map h_l as $(1+\gamma_l) \odot h_l + \beta_l$. (b) Inference: pseudo masks sampled from a mask-only diffusion model trained on the limited real masks are used to guide the synthesis of new image–mask pairs. (c) Downstream use: the synthesized pairs are combined with the limited real training set to augment downstream UNet++ segmentation, and performance is evaluated on the held-out real test set.

2. Related Work

Synthetic Labeled Data for Medical Image Segmentation. Synthetic labeled data has been widely used to mitigate the scarcity of pixel-level annotations in medical image segmentation. Earlier work explored conventional augmentation and mask-guided generative pipelines [13, 15], while recent diffusion-based methods have shown stronger potential for segmentation-oriented data augmentation under low-data settings [3, 6, 7, 10–12, 17]. These studies motivate synthetic supervision as a practical strategy for downstream medical segmentation.

Mask-Conditioned and Paired Diffusion Generation. Spatial conditioning has been widely used to improve controllability in medical image synthesis. *SegGuidedDiff* generates anatomically controllable images from mask layouts [4], while *GenSeg* and related methods explore mask-conditioned image generation or paired synthetic supervision for segmentation [5, 16, 17]. Conditional modulation mechanisms such as FiLM and SPADE further provide effective ways to inject structural or semantic guidance [8, 9]. In contrast, we focus on structure-consistent joint image–mask synthesis for data-scarce medical segmentation, using an additional mask-conditioned control branch to strengthen spatial correspondence between generated images and masks.

3. Method

3.1. Overview

Let $\mathcal{D} = \{(x_i, m_i)\}_{i=1}^{N}$ denote the limited set of real training pairs, where $x_i \in \mathbb{R}^{C \times H \times W}$ is a medical image and $m_i \in \{0,1\}^{1 \times H \times W}$ is its corresponding binary mask. Instead of generating only an image conditioned on a mask, we formulate synthetic data generation as *joint image–mask synthesis*. Specifically, we define the joint target

$$y_i = [x_i, m_i] \in \mathbb{R}^{(C+1) \times H \times W}, \tag{1}$$

where $[\cdot, \cdot]$ denotes channel-wise concatenation. Gaussian noise with the same shape as $y = [x, m]$ is added at each diffusion step, and all channels are modeled jointly within a single diffusion process.

We train a conditional diffusion model to learn the paired distribution $p_\theta(y \mid c)$, where $c \in \mathbb{R}^{1 \times H \times W}$ is an additional conditioning mask. The model jointly synthesizes the image and target mask, while an additional mask-conditioned FiLM control branch provides structural guidance to improve consistency. Training follows the standard diffusion noise-prediction objective.

3.2. Joint Image–Mask Synthesis with Mask-Conditioned FiLM Control

A key design of our method is to distinguish the *target mask* from the *conditioning mask*. The target mask m is part of the output pair $y = [x, m]$ and is jointly synthesized with the image, whereas the conditioning mask c is an external control signal that guides generation toward a desired lesion structure. Although both are derived from the same ground-truth annotation during training, they serve different roles.

To inject structural guidance, we adopt FiLM. FiLM provides a lightweight and flexible way to inject mask-dependent feature modulation at multiple resolutions, which matches our goal of using the conditioning mask as an auxiliary structural guidance signal rather than as the sole

generative target. Let $h_l \in \mathbb{R}^{C_l \times H_l \times W_l}$ denote an intermediate feature map at network level l and $c_l \in \mathbb{R}^{1 \times H_l \times W_l}$ denote the conditioning mask resized to the same spatial resolution as h_l. c_l is obtained by progressively downsampling the input conditioning mask along the encoder path. A learnable 1×1 convolution $G_l(\cdot)$ maps c_l to spatial FiLM parameters:

$$[\gamma_l, \beta_l] = G_l(c_l), \tag{2}$$

where $\gamma_l, \beta_l \in \mathbb{R}^{C_l \times H_l \times W_l}$ are the modulation tensors.

The FiLM-modulated feature is then written as

$$\mathrm{FiLM}(h_l; c_l) = (1 + \gamma_l) \odot h_l + \beta_l, \tag{3}$$

where $\odot$ denotes element-wise multiplication. Intuitively, γ_l adjusts the local feature magnitude according to the conditioning mask, while β_l provides an additive spatial bias that further steers the representation toward the desired lesion layout.

In our implementation, FiLM is applied at each downsampling stage and at the bottleneck of the U-Net, using conditioning masks at the corresponding resolutions. The modulation is combined with standard timestep conditioning inside the residual blocks. We do not apply FiLM in the upsampling path, which keeps the control branch lightweight while preserving strong structural guidance in the encoder and bottleneck.

3.3. Training and Inference

During training, the conditioning mask is taken from the same ground-truth annotation as the target mask: the latter is part of the joint target $y = [x, m]$, while the former is separately provided to the FiLM branch for structural guidance. To enable conditional sampling under the same low-data setting, we train a separate mask-only diffusion model using masks from the same training pairs. This model is a single-channel, unconditional variant of the same diffusion framework and is used to generate pseudo masks for structural conditioning. During inference, a pseudo mask sampled from the mask-only diffusion model is fed to the FiLM branch, and the joint diffusion model generates a synthetic image–mask pair conditioned on it. The resulting synthetic pairs are then combined with the limited real training set for downstream UNet++ augmentation.

4. Experiments

4.1. Experimental Setup

Datasets and baselines. We evaluate our method on two medical imaging datasets: HyperKvasir [2] and JSRT [14]. We compare with two representative synthetic-data baselines, *GenSeg* [16] and *SegGuidedDiff* [4], as well as ablated variants of our own model.

Low-data protocol. All experiments are conducted under a strict low-data setting using only 32 ground-truth (GT) image–mask pairs for training. These 32 pairs are selected from the original training set using a greedy diversity-based selection strategy, with the goal of choosing samples that are as dissimilar as possible and therefore cover broader structural variation. The same 32 GT pairs are used consistently throughout the entire pipeline: they are used to train the main joint image–mask generation model, to train the separate mask-only diffusion model on the corresponding 32 GT masks for pseudo-mask generation, and to train the downstream UNet++ segmentation model. During inference, pseudo masks sampled from the mask-only diffusion model are used as conditioning masks.

Table 1. Downstream UNet++ segmentation results under the 32-pair low-data protocol. We report GT-only training and augmentation with 128 generated image–mask pairs (32GT+128GEN) on two representative datasets, HyperKvasir and JSRT. Best results are shown in bold.

Method	HyperKvasir mIoU ↑	HyperKvasir Dice ↑	JSRT mIoU ↑	JSRT Dice ↑
32GT	0.5702	0.6872	0.9443	0.9713
GenSeg [16]	0.6065	0.7159	**0.9501**	**0.9743**
SegGuidedDiff [4]	0.5983	0.7105	0.9454	0.9718
Our Method	**0.6237**	**0.7319**	0.9494	0.9740

Downstream evaluation. Generated image–mask pairs are combined with the same 32-pair real training set to augment downstream UNet++ training. We compare *GT-only* and *GT+N generated samples* under the same protocol for all methods, with $N \in \{32, 64, 128, 256\}$. For each synthetic-data budget N, we construct 10 distinct synthetic subsets by sampling N pairs from the generated pool using different fixed random seeds, and train the downstream UNet++ model once on each subset. Final results are reported as the average test performance over these 10 runs. The main paper reports *32GT* and the representative *32GT+128GEN* setting, with the complete synthetic-data budget sweep provided in the supplementary material.

Implementation details. All generative models were trained on a single NVIDIA RTX A5000 GPU (24 GB memory) using Adam with learning rate 1×10^{-4} and $\beta = (0.9, 0.99)$. We used batch size 16, fp16 training, image size 128×128, 1000 diffusion timesteps, 500 sampling timesteps, cosine noise schedule, and EMA decay 0.999.

4.2. Main Results

Table 1 shows the downstream segmentation results on *HyperKvasir* and *JSRT*. On *HyperKvasir*, an RGB image–mask setting with clear room beyond the 32GT baseline, all three generation-based methods improve over GT-only training, and our method performs best. In contrast, *JSRT* is a near-ceiling grayscale setting, where all three methods remain highly competitive and the performance gaps are

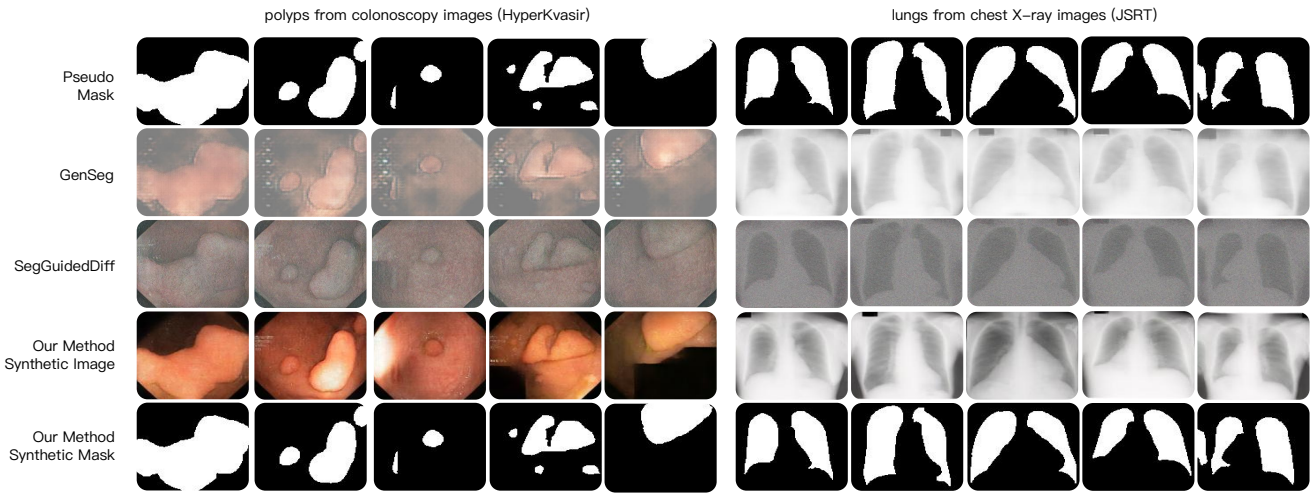

Figure 2. Representative qualitative results on *HyperKvasir* and *JSRT* under the 32-pair low-data setting. For each dataset, we show the pseudo conditioning mask, images generated by *GenSeg* and *SegGuidedDiff*, and the image–mask pair generated by our method.

marginal. GenSeg achieves slightly higher scores on *JSRT* in some mixed real–synthetic settings, but these small differences should not be over-interpreted. One possible reason is that GenSeg's direct mask-to-image paradigm is already sufficient when the conditional lung mask largely determines the target structure, whereas our joint image–mask formulation shows clearer benefits in the more challenging *HyperKvasir* setting.

4.3. Qualitative Results

Figure 2 shows representative pseudo conditioning masks, generated images, and generated masks on *HyperKvasir* and *JSRT*. *HyperKvasir* corresponds to a more challenging RGB image–mask setting, where our method produces structurally aligned synthetic image–mask pairs, consistent with its stronger quantitative improvement. In contrast, *JSRT* is a simpler grayscale setting, where the generated structures are visually more regular and all methods produce high-quality results, matching the near-ceiling quantitative performance.

4.4. Ablation Study

We ablate two design choices in our framework on HyperKvasir: joint image–mask modeling and FiLM-based mask-conditioned control. Specifically, Table 2 compares *Joint image–mask + FiLM*, *Joint image–mask only*, and *Image-only + FiLM*. The results show that the joint image–mask variants perform substantially better than *Image-only + FiLM*, indicating that explicit paired modeling is necessary and that FiLM conditioning alone is insufficient. Among the joint variants, *Joint image–mask + FiLM* performs best, showing that the additional FiLM-based control branch is most effective when combined with joint image–mask synthesis and further improves the utility of the generated pairs for downstream segmentation.

Table 2. Ablation on HyperKvasir under the 32-pair low-data protocol. Joint denotes joint image–mask synthesis, and FiLM denotes mask-conditioned control. We report 32GT and 32GT+128GEN.

Setting	Joint	FiLM	mIoU ↑	Dice ↑
32GT	–	–	0.5702	0.6872
32GT+128GEN	✓	✓	**0.6237**	**0.7319**
32GT+128GEN	✓	–	0.6042	0.7160
32GT+128GEN	–	✓	0.4220	0.5502

5. Conclusion

We presented a structure-consistent framework for joint image–mask synthesis in data-scarce medical image segmentation. By combining joint paired modeling with a lightweight FiLM-based mask-conditioned control branch, the proposed method generates coherent synthetic image–mask pairs and yields strong downstream UNet++ performance under the same low-data protocol, especially in the more challenging RGB setting. Ablation results show that explicit joint modeling is necessary, and that FiLM-based control is most effective when built on top of joint synthesis. One limitation is that the diversity and quality of generated pairs are still bounded by the limited real masks used to train the pseudo-mask generator, which can lead to saturation when the synthetic-data budget becomes large. Future work may explore more adaptive conditioning-mask generation, quality-aware synthetic sample selection, and real–synthetic weighting strategies for more robust low-data segmentation augmentation.

6. Acknowledgement

This work was supported in part by U.S. NSF grant DBI-2422619.

References

[1] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, pages 1501–1510, 2017. 6

[2] 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*, pages 451–462, 2020. 3

[3] Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. Diffusion models in medical imaging: A comprehensive survey. *Medical Image Analysis*, 88:102846, 2023. 1, 2

[4] Nicholas Konz, Yuwen Chen, Haoyu Dong, and Maciej A. Mazurowski. Anatomically-controllable medical image generation with segmentation-guided diffusion models. In *MICCAI*, 2024. 1, 2, 3

[5] Jiawei Mao, Yuhan Wang, Yucheng Tang, Daguang Xu, Kang Wang, Yang Yang, Zongwei Zhou, and Yuyin Zhou. Medsegfactory: Text-guided generation of medical image-mask pairs. In *ICCV*, pages 21525–21535, 2025. 1, 2

[6] Pedro A. Morão, Yasna Forghani, Nuno Loução, Pedro Gouveia, Mario A. T. Figueiredo, and Joao Santinha. Data augmentation for medical imaging: Counterfactual simulation of acquisition parameters via conditional diffusion model. In *Medical Imaging with Deep Learning*, 2025. 1, 2

[7] Maham Nazir, Muhammad Aqeel, and Francesco Setti. Diffusion-based data augmentation for medical image segmentation. In *ICCVW*, pages 1341–1350, 2025. 1, 2

[8] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, pages 2332–2341, 2019. 2, 6

[9] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: visual reasoning with a general conditioning layer. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*. AAAI Press, 2018. 1, 2

[10] Sivaramakrishnan Rajaraman, Zhaohui Liang, Zhiyun Xue, and Sameer Antani. Addressing class imbalance with latent diffusion-based data augmentation for improving disease classification in pediatric chest x-rays. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 5059–5066, 2024. 1, 2

[11] Daniel G. Saragih, Atsuhiro Hibi, and Pascal N. Tyrrell. Using diffusion models to generate synthetic labeled data for medical image segmentation. *International Journal of Computer Assisted Radiology and Surgery*, 19(8):1615–1625, 2024.

[12] Sumin Seo, In Kyu Lee, Hyun-Woo Kim, Jaesik Min, and Chung-Hwan Jung. Diffusion-based user-guided data augmentation for coronary stenosis detection. In *MICCAI*. Springer Nature Switzerland, 2025. 1, 2

[13] Younghak Shin, Hemin Ali Qadir, and Ilangko Balasingham. Abnormal colon polyp image synthesis using conditional adversarial networks for improved detection performance. *IEEE Access*, 6:56007–56017, 2018. 2

[14] Junji Shiraishi, Shigehiko Katsuragawa, Junpei Ikezoe, Tsuneo Matsumoto, Takeshi Kobayashi, Ken-ichi Komatsu, Mitate Matsui, Hiroshi Fujita, Yoshie Kodera, and Kunio Doi. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *American journal of roentgenology*, 174 (1):71–74, 2000. 3

[15] Vajira Thambawita, Pegah Salehi, Sajad Amouei Sheshkal, Steven A. Hicks, Hugo L. Hammer, Sravanthi Parasa, Thomas de Lange, Pål Halvorsen, and Michael A. Riegler. Singan-seg: Synthetic training data generation for medical image segmentation. *PLOS ONE*, 17(5):1–24, 2022. 2

[16] Shanshan Wang, Cheng Li, Rongpin Wang, Zaiyi Liu, Meiyun Wang, Hongna Tan, Yaping Wu, Xinfeng Liu, Hui Sun, Rui Yang, Xin Liu, Jie Chen, Huihui Zhou, Ismail Ben Ayed, and Hairong Zheng. Annotation-efficient deep learning for automatic medical image segmentation. *Nature Communications*, 12(1):5915, 2021. 1, 2, 3

[17] Li Zhang, Basu Jindal, Ahmed Alaa, Robert Weinreb, David Wilson, Eran Segal, James Zou, and Pengtao Xie. Generative ai enables medical image segmentation in ultra low-data regimes. *Nature Communications*, 16(1):6486, 2025. 1, 2

[18] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *TMI*, 2019. 1