Noise-Consistent Siamese-Diffusion for Medical Image Synthesis and Segmentation

Supplementary Material

7. Additional Explanation of Architecture

In this section, we provide additional preliminary knowledge explanations for Sec. 3 and present a detailed training algorithm of *Siamese-Diffusion*.

7.1. Architecture of Siamese-Diffusion

Numerous methods [31, 56, 59] based on pre-trained Stable Diffusion [36] achieve fine control over synthesized images using external HyperNetworks [17] to process structural inputs like segmentation masks, depth maps, and sketches. T2I-Adapter [31] enables control with a lightweight feature extractor without updating Stable Diffusion's parameters. Meanwhile, ControlNet [59] demonstrates that precise control over synthesized images is achievable regardless of whether the denoising U-Net decoder parameters are updated. The key to the success of these methods lies in ensuring the effective fusion of prior control features and noisy image features in the latent space.

Although ControlNet [59] demonstrates that updating the denoising U-Net decoder parameters is not essential for achieving fine control, such updates can improve the quality of synthesized images. In our proposed *Siamese-Diffusion* framework, *Mask-Diffusion* leverages collaborative updates from both the external network (comprising the cascaded Dense Hint Input module and ControlNet) and the denoising U-Net decoder parameters. This collaboration ensures the effective fusion of mask features c_m with noisy image features z_t in the latent space, enabling *Mask-Diffusion* to operate independently during sampling.

Image-Diffusion relies exclusively on the external feature extraction network for unidirectional fusion with noisy image features in the latent space. By leveraging *Noise Consistency Loss*, the U-Net decoder parameters are adjusted, enhancing the fusion between mask-image joint prior control features c_{mix} and noisy image features z_t. As a result, *Image-Diffusion*, benefiting from the added image prior control, achieves more accurate noise predictions compared to *Mask-Diffusion*. *Noise Consistency Loss* further propagates these benefits to *Mask-Diffusion*, refining its parameters and enabling it to independently synthesize images with enhanced morphological characteristics. The "copy" operation in Fig. 2(a) and the shared c_m in Eq. (5) ensure that the differences in noise predictions between the two processes are solely due to the additional image prior control, stabilizing the propagation process.

Mask-Diffusion and *Image-Diffusion* share a unified diffusion model. The Siamese architecture stands out from traditional distillation models [19, 30], which require separate training of a teacher network with mask-image prior control and a student network with mask-only prior control. Thus, the Siamese structure relies on fewer parameters, making it more resource efficient for training.

Algorithm 1 Training algorithm of Siamese-Diffusion.

```
for Every batch of size N do
    for $(x_0, y_0)$ in this batch do
        Sample:
        $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), t \sim \mathcal{U}(\{0, 1, ..., T\})$
        Encode Image into Latent Space:
        $z_0 = \mathcal{E}(x_0)$
        Encode Prior Controls into Latent Space:
        $c_i = \mathcal{F}(x_0), c_m = \mathcal{F}(y_0)$
        Mix Mask and Image Prior Controls:
        $c_{mix} = w_i \cdot c_i + w_m \cdot sg[c_m]$
        Noise Image in the Latent Space:
        $z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon$
        Calculate Mask Denoising Loss:
        $\mathcal{L}_m = \|\epsilon_\theta(z_t, c_m) - \epsilon\|^2$
        Copy the Parameter of Mask-Diffusion:
        $\theta' = \text{DeepCopy}(\theta)$
        Calculate Image Denoising Loss:
        $\mathcal{L}_i = \|\epsilon_{\theta'}(z_t, c_{mix}) - \epsilon\|^2$
        Calculate Noise Consistency Loss:
        $\mathcal{L}_c = w_c \cdot \|\epsilon_\theta(z_t, c_m) - sg[\epsilon_{\theta'}(z_t, c_{mix})]\|^2$
        Single-Step Sampling:
        $z'_0 = \frac{z_t - \sqrt{1 - \bar{\alpha}_t} sg[\epsilon_{\theta'}(z_t, c_{mix})]}{\sqrt{\bar{\alpha}_t}}$
        Noise Image in the Latent Space:
        $z'_t = \sqrt{\bar{\alpha}_t} z'_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon$
        Calculate Augmented Mask Denoising Loss:
        $\mathcal{L}_{m'} = w_a \cdot \|\epsilon_\theta(z'_t, c_m) - \epsilon\|^2$
        Update the Parameters:
        $\mathcal{L} = \mathcal{L}_m + \mathcal{L}_i + \mathcal{L}_c + \mathcal{L}_{m'}$
    end for
end for
```

7.2. Dense Hint Input Module

As mentioned above, simultaneously updating the denoising U-Net parameters enhances the quality of synthesized images. However, *Image-Diffusion* compromises the fusion between image prior control features c_i and noisy image features z_t. In [56], high-density semantic image features achieve unidirectional fusion with noisy image features using a robust external network, without requiring collaborative updates to the denoising U-Net parameters. To enhance the fusion between mask-image joint prior control features c_{mix} and noisy image features z_t, we substitute the original sparse Hint Input module with a more powerful Dense Hint Input module. Furthermore, [56] demonstrates that a stronger external HyperNetwork can effectively accommodate low-density semantic inputs (*e.g.*, segmentation masks). As a result, *Mask-Diffusion* and *Image-Diffusion* share a unified external feature extractor in our approach.

7.3. Online-Augmentation

Typically, due to the low accuracy of noise prediction in single-step sampling, diffusion models (*e.g.*, [21, 44, 45]) rely on Predict-Correct (PC) multi-step sampling. However, using multi-step sampling during training is both time-consuming and memory-intensive, as it requires storing gradient information [27], which limits the feasibility of generating new sample pairs online to augment the training set. By leveraging the advantages of the *Siamese-Diffusion* training paradigm, single-step sampling can be used for online augmentation in *Mask-Diffusion*.

7.4. Training Algorithm of Siamese-Diffusion

The detailed training algorithm of *Siamese-Diffusion* is presented in Algorithm 1. For simplicity, the prompt c_t and timestep t are omitted. $\mathcal{E}$ represents the encoder of VQ-VAE [50], whose parameters are frozen. $\mathcal{F}$ represents the external feature extraction network, consisting of the cascaded Dense Hint Input module and ControlNet, which are jointly trained with the Diffusion model.

8. Test Set for Segmentation Models

This section provides detailed descriptions of the test sets used to evaluate segmentation models across various datasets, as discussed in Sec. 4.1.

Polyps Dataset: Following [14], evaluations are conducted on five public polyp datasets: EndoScene [51] (60 samples), CVC-ClinicDB/CVC-612 [2] (62 samples), Kvasir [26] (100 samples), CVC-ColonDB [46] (380 samples), and ETIS [42] (196 samples). *Overall* represents the weighted average results of these five test sets.

ISIC2016 & ISIC2018 Datasets: Evaluations are performed on their official test sets.

Stain & Feces Datasets: As described in the main paper, the Stain dataset (500 samples) and the Faeces dataset (458 samples) are divided into training, validation and test sets in a 3:1:1 ratio, resulting in test sets of 100 and 92 samples.

9. Image Quality Comparison

This section provides additional quantitative and qualitative analysis of image quality assessment results on other datasets, as discussed in Sec. 5.1.

Table 7. Comparison of synthetic skin lesion (ISIC2016) image quality generated by each respective mask-only method, evaluated using FID [18], KID [3], CLIP-I [38], LPIPS [60], CMMD [25] and MOS metrics.

Methods	FID (↓)	KID (↓)	CLIP-I (↑)	LPIPS (↓)	CMMD (↓)	MOS (↑)
T2I-Adapter [31]	234.474	0.1912	0.774	0.688	2.733	-
ControlNet [59]	68.327	**0.0267**	0.820	0.667	**0.688**	1.68
Ours	**64.208**	0.0299	**0.827**	**0.657**	0.733	**1.96**

Table 8. Comparison of synthetic skin lesion (ISIC2018) image quality generated by each respective mask-only method, evaluated using FID [18], KID [3], CLIP-I [38], LPIPS [60], CMMD [25] and MOS metrics.

Methods	FID (↓)	KID (↓)	CLIP-I (↑)	LPIPS (↓)	CMMD (↓)	MOS (↑)
T2I-Adapter [31]	224.446	0.1751	0.796	**0.672**	2.621	-
ControlNet [59]	45.490	0.0286	**0.809**	**0.672**	0.794	1.22
Ours	**44.036**	**0.0258**	0.808	0.673	**0.701**	**2.12**

9.1. Quantitative Evaluation

For the five datasets, we use various image quality evaluation metrics, including Fréchet Inception Distance (FID) [18], Kernel Inception Distance (KID) [3], CLIP-Image (CLIP-I) [38], Learned Perceptual Image Patch Similarity (LPIPS) [60] and CLIP-Maximum Mean Discrepancy (CMMD) [25]. Additionally, for medical datasets, we employ the Mean Opinion Score (MOS), calculated by averaging experienced clinicians' ratings of synthetic image quality. Distinct evaluation standards are applied to the Polyps and Skin Lesion datasets based on clinician suggestions

Polyps Dataset: As shown in Tab. 1, the commonly used non-human evaluation metrics have been discussed in the main paper. Here, we focus on the MOS metric. Three professional clinicians assess the quality of polyp images generated by SinGAN-Seg [47], ControlNet [59], our *Siamese-Diffusion*, and real data. T2I-Adapter [31] and ArSDM [14] are excluded based on clinicians' suggestions due to their unrealistic results. To ensure the reliability of the evaluation results and minimize the fatigue of the clinicians, 50 image groups are randomly selected, each containing images synthesized using the same mask by the four methods (including real images), totaling 200 polyp images. Clinicians view one image at a time, scoring it as "Real" (1 point) or "Synthetic" (0 point) and providing a confidence score ranging from 1 to 10. Our method achieves the highest MOS score of 0.587 with a confidence level of 6.04, demonstrating superior quality in synthesized polyp images. 50 real images are used to assess preference, yielding MOS score of 0.9 and confidence level of 6.17.

ISIC2016 Dataset: As shown in Tab. 7, we compare the synthesized skin lesion image quality of T2I-Adapter [31], ControlNet [59], and our *Siamese-Diffusion*. Although KID [3] and CMMD [25] scores are slightly lower than those of ControlNet [59], our method achieves the best scores in FID [18], CLIP-I [38], LPIPS [60], and MOS, demonstrating superior overall performance. About MOS evaluation, one experienced clinician assesses the quality of skin lesion images generated by ControlNet [59], our *Siamese-Diffusion*, and real data. T2I-Adapter [31] is excluded for the same reason as in the Polyps dataset. Considering potential biases in single-evaluator assessments, these results are presented as reference points rather than definitive benchmarks. Following clinician suggestions, 50 image groups are randomly selected, each containing images synthesized using the same mask by the three methods (including real images), totaling 150 skin lesion images. The clinician views three images at a time and assigns rankings from 1 to 3 (highest). Our method achieves the highest MOS score of 1.96. 50 real images are used to assess preference, yielding an MOS score of 2.36.

ISIC2018 Dataset: As shown in Tab. 8, similar to the ISIC2016 dataset, our method performs better overall. The ISIC2018 dataset ($2,594$ samples) contains significantly more trainable data compared to the ISIC2016 dataset (900 samples), resulting in markedly improved FID [18] and KID [3] scores. This observation supports the notion that increasing the amount of training data for generative models can improve the quality of synthesized images. Incredibly, T2I-Adapter [31] and ControlNet [59] achieve identical LPIPS [60] scores of 0.672, which are marginally better than the 0.673 achieved by our *Siamese-Diffusion*. However, the visualization results in Fig. 10(b) contradict this outcome, suggesting that LPIPS [60] may not reliably reflect alignment with human perception. The MOS evaluation standards applied to the ISIC2018 dataset are the same as those used for ISIC2016. Our method achieves the highest MOS score of 2.12. Additionally, 50 real images are used to assess preference, yielding an MOS score of 2.66.

Stain & Faeces Datasets: As shown in Tab. 9 and Tab. 10, we compare the synthesized image quality of DFMGAN [15], AnomalyDiffusion [23], T2I-Adapter [31], ControlNet [59], and our *Siamese-Diffusion*. Our method outperforms the others on both datasets, demonstrating the superiority of our approach. Shockingly, T2I-Adapter [31] achieves the best CMMD [25] score on the Faeces dataset, but discrepancies with the visualization in Fig. 11(c) highlight potential limitations of the CMMD metric.

9.2. Qualitative Evaluation

In this section, we provide a qualitative analysis of synthesized images from four datasets generated by different methods. For sensitivity reasons, synthesized images for the Faeces dataset are not displayed.

Polyps Dataset: Fig. 8 presents visualizations of polyp images synthesized by different generative models. The differences between each method have been discussed in the main paper. Additional examples further demonstrate that our method produces images with rich morphological characteristics, validating the superiority of our approach. Notably, the "editing-like" approach of SinGAN-Seg [47] generates minimal artifacts when the mask varies slightly, aligning with human perception when viewed without zooming in. However, when the mask undergoes significant variations, the artifacts become extremely pronounced, undermining the realism of the synthesized images.

Table 9. Comparison of synthetic stain image quality generated by each respective mask-only method, evaluated using FID [18], KID [3], CLIP-I [38], LPIPS [60], and CMMD [25] metrics.

Methods	FID (↓)	KID (↓)	CLIP-I (↑)	LPIPS (↓)	CMMD (↓)
DFMGAN [47]	242.780	0.1619	0.712	0.781	3.733
AnomalyDiffusion [59]	165.732	0.0791	0.763	0.778	1.296
T2I-Adapter [31]	209.260	0.1371	0.765	0.778	1.296
ControlNet [14]	123.818	0.0298	0.769	0.731	1.213
Ours	**115.546**	**0.0206**	**0.773**	**0.719**	**1.183**

Table 10. Comparison of synthetic faece image quality generated by each respective mask-only method, evaluated using FID [18], KID [3], CLIP-I [38], LPIPS [60], and CMMD [25] metrics.

Methods	FID (↓)	KID (↓)	CLIP-I (↑)	LPIPS (↓)	CMMD (↓)
DFMGAN [47]	299.032	0.2156	0.639	0.760	6.369
AnomalyDiffusion [59]	220.003	0.1181	0.733	0.754	2.232
T2I-Adapter [31]	207.814	0.1118	0.778	0.651	**1.264**
ControlNet [14]	166.567	0.0843	0.765	0.651	1.701
Ours	**143.736**	**0.0485**	**0.786**	**0.643**	1.337

ISIC2016 & ISIC2018 Datasets: Fig. 9 and Fig. 10 present visualizations of skin lesion images generated by various models, revealing phenomena similar to those observed in the Polyps dataset. T2I-Adapter [31] generates unrealistic images with a uniform "style". Compared to ControlNet [59], our method demonstrates superior performance in mask alignment, morphological texture, and color, validating the effectiveness and superiority of our approach.

Stain Dataset: Fig. 11 presents visualizations of stain images synthesized by various models. DFMGAN [15] cannot control the synthesis with the specified masks and performs poorly when data is scarce. AnomalyDiffusion [23] demonstrates poor alignment of the mask and stain, especially when the mask area is small. T2I-Adapter [31] generates images with a unified background "style" and exhibits low content density. In terms of content density, our method outperforms ControlNet [59], generating richer content, which corresponds to the richness of morphological characteristics in medical images, thus validating the effectiveness and superiority of our approach.

10. Qualitative Analysis of Each Component

In this section, we present additional images illustrating the impact of each component, as shown in Fig. 12, to further substantiate the conclusions drawn in Sec. 5.3.1.

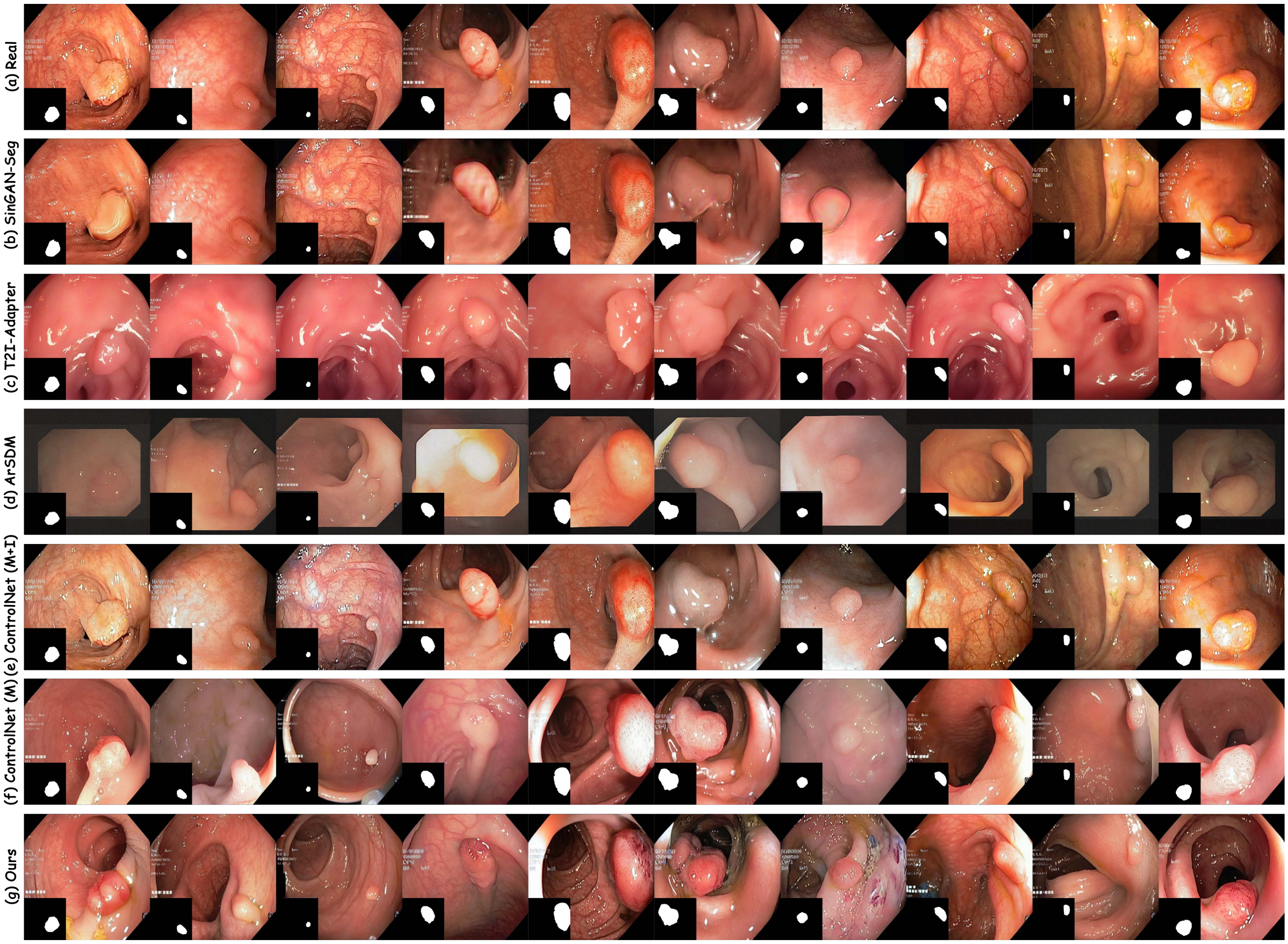

Figure 8. (a) Examples of real polyp images. (b)–(g) Examples of synthetic polyp images generated by each respective method. “M” denotes that mask-only prior control, while “M+I” denotes mask-image joint prior control. The synthetic polyp images generated by our method achieve competitive fidelity while also exhibiting diversity (Zoom in for better visualization).

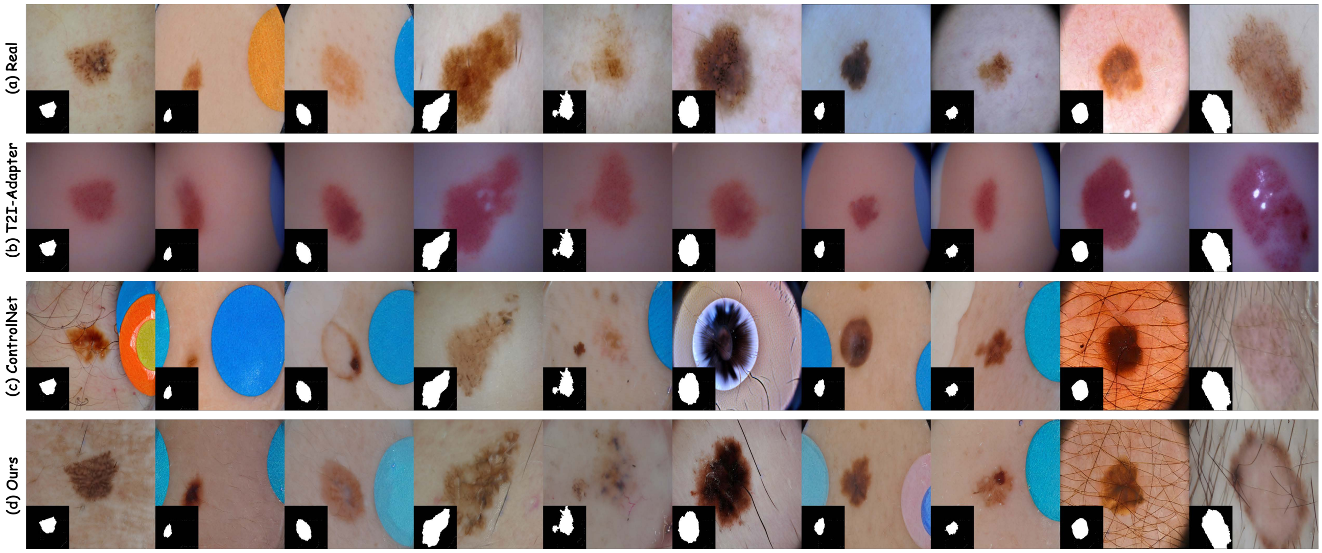

Figure 9. (a) Examples of real skin lesion (ISIC2016) images. (b)–(d) Examples of synthetic skin lesion images generated by each respective method. The synthetic skin lesion images generated by our method achieve competitive fidelity while also exhibiting diversity (Zoom in for better visualization).

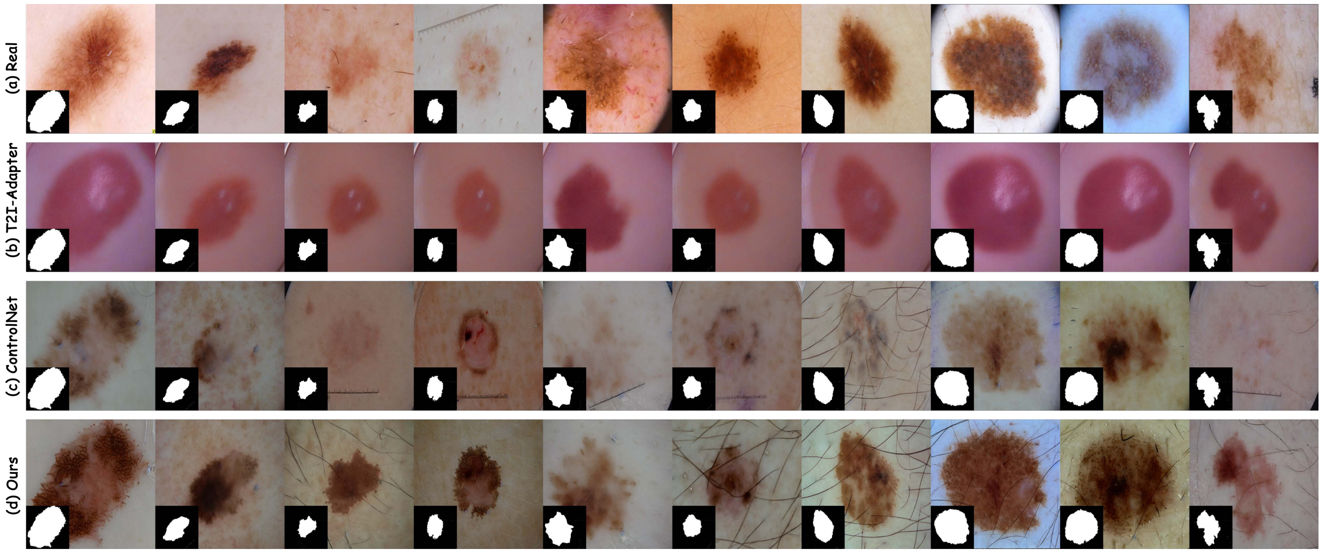

Figure 10. (a) Examples of real skin lesion (ISIC2018) images. (b)–(d) Examples of synthetic skin lesion images generated by each respective method. The synthetic skin lesion images generated by our method achieve competitive fidelity while also exhibiting diversity (Zoom in for better visualization).

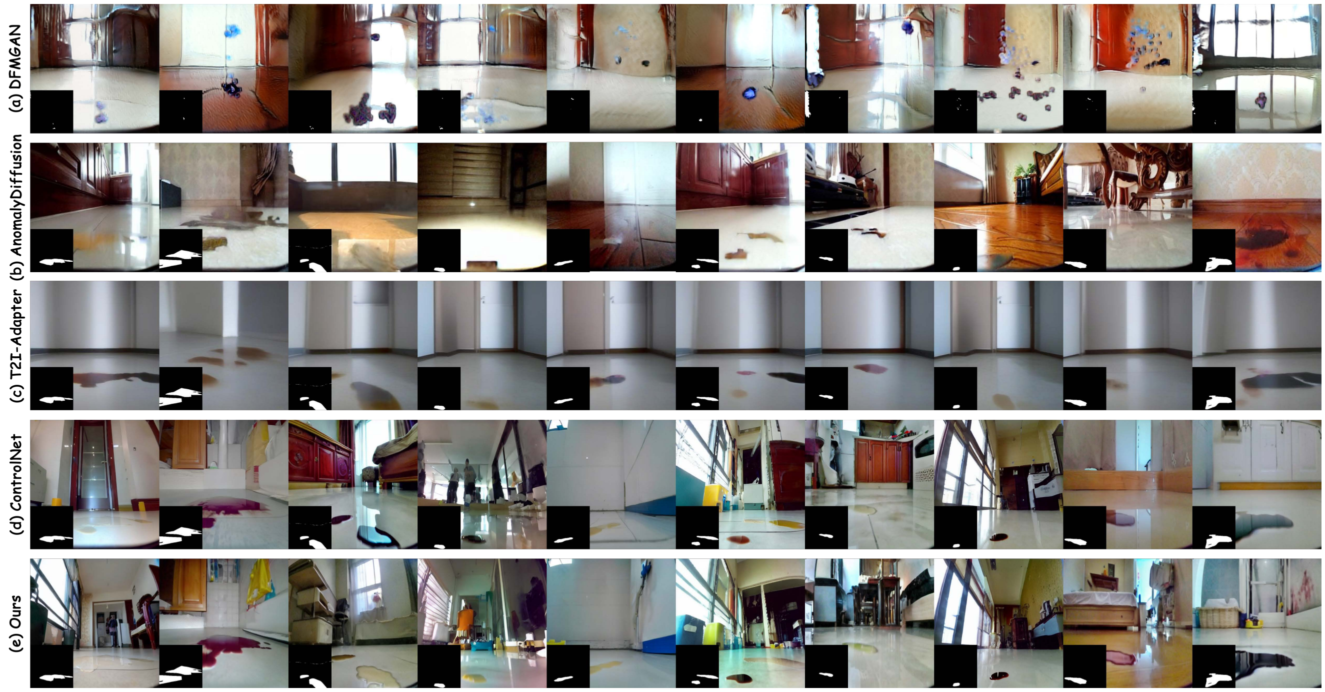

Figure 11. (a)–(e) Examples of synthetic stain images generated by each respective method are shown. DFMGAN [15] cannot control the synthesis with the specified masks. The synthetic stain images generated by our method exhibit higher semantic density, meaning richer content, which corresponds to the rich morphological characteristics in medical images. Additionally, these synthetic stain images achieve competitive fidelity while also demonstrating diversity (Zoom in for better visualization).

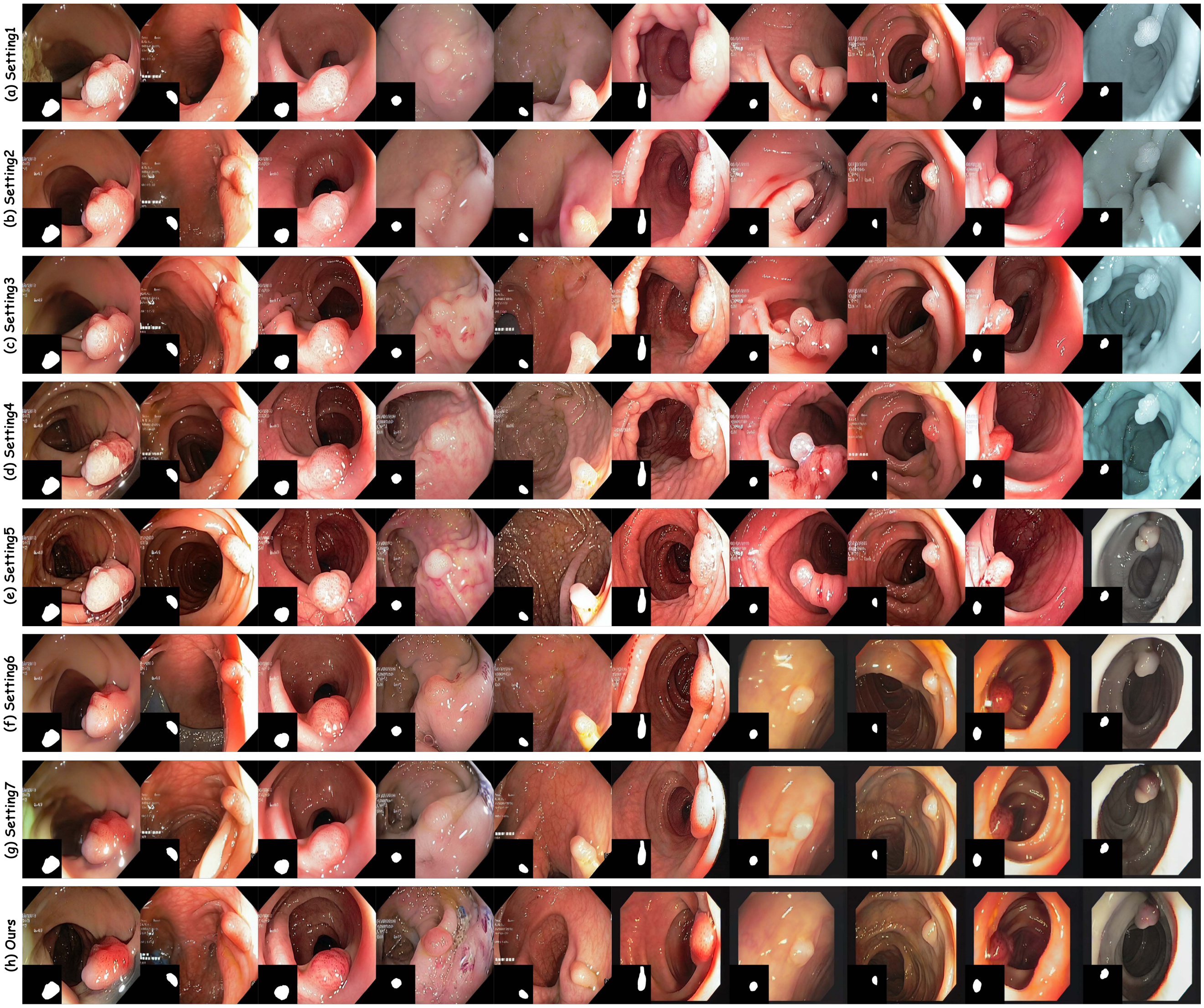

Figure 12. Visualizing the impact of different components on the synthesis of polyp images (Zoom in for better visualization).