PathDiff: Histopathology Image Synthesis with Unpaired Text and Mask Conditions

Supplementary Material

1. Scaling Augmentation in Downstream Tasks

To evaluate the impact of synthetic data augmentation on downstream tasks, we designed three augmented sets for training the CellViT [9] model on the PanNuke [4] dataset. These augmented datasets were generated from the training split, conditioned on mask, and evaluated on the real test split. They were incrementally added to the training process while keeping the size of the real train split constant.

Scaling the augmentation set progressively improves the classification and segmentation performance of the training data, as shown in Fig. 1, with the 3x and 2.5x augmented synthetic datasets outperforming the relatively smaller ones on all metrics. After 2.5x, the performance metrics plateau. These results demonstrate that PathDiff effectively contributes valuable synthetic data in every augmented set, consistently improving model performance across all downstream tasks as the size of the synthetic set increases, highlighting PathDiff generates diverse high-quality data for histopathology image analysis.

2. Qualitative Comparison of Synthetic Images

In this section, we present a qualitative comparison of synthesized images generated by PathDiff, DiffMix [12], SDM [17], and ControlNet [19].

2.1. Mask-to-Image examples

Fig. 2 shows a comparison of synthetic images generated on the PanNuke [4], CoNIC [5], and MoNuSAC [16] datasets. As illustrated in the figure, images generated by DiffMix appear very coarse with additional artifacts and fail to preserve the accurate stain colors observed in the original images. Consistent with observations reported by [11], we find that SDM-generated images display unrealistic color overlay artifacts. The color distribution of ControlNet-generated images appears highly inconsistent, being significantly inaccurate in some cases while better than others in certain instances. On the other hand, PathDiff accurately follows the cell mask and maintains the cell stain colors as seen in the original images.

2.2. Text-to-Image examples

Fig. 4 shows samples generated by PathDiff and ControlNet. As with images conditioned on masks, ControlNet fails to preserve details in the original image and exhibits implausible colors uncommon in histopathology images. This explains the high FID and KID values compared to PathDiff in Tab.3 of the main paper.

2.3. Unified Paired Conditions Sampling

Fig. 3 shows images generated from paired Text and silver standard masks. PathDiff generated images incorporated guidance from both Text and Mask successfully and look significantly better than ControlNet.

3. Domain Expert Assessment

We acknowledge that traditional fidelity metrics like FID [7] are only somewhat applicable to histological images as large image datasets like ImageNet [2] are unlikely to contain images from this specific domain. Therefore, we conduct expert evaluation to validate the efficiency of the generated samples. We surveyed two domain experts—a physician and a pathology researcher—to review the generated data and assess if the samples accurately reflect the characteristics of real specimens.

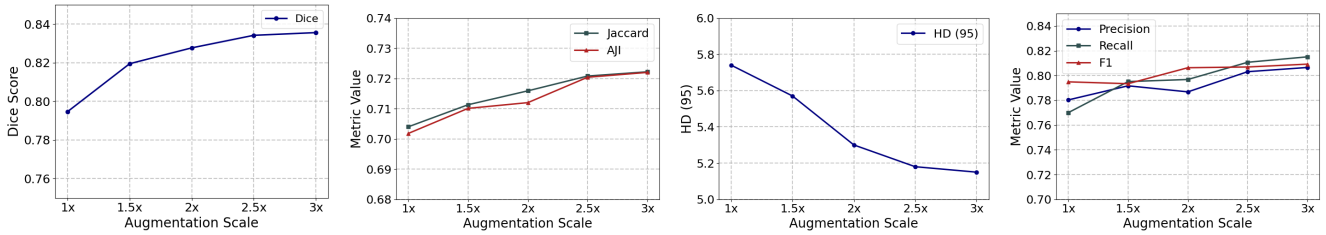
Image Preference Experiment: We presented domain experts with a total of 200 synthetic images (Quadruplets of 50) generated from PathDiff, SDM [17], ControlNet [19], DiffMix [12]. Each Quadruplet of images was generated using the same conditional mask. Domain experts were asked to choose one of the four images that looked most real. As shown in Fig. 5, both domain experts preferred PathDiff-generated images significantly more than the existing SOTA methods, indicating our generated images look more realistic to an expert eye compared to others.

	Predicted Label	
Actual Label	**Real**	**Synthetic**
Synthetic	15	12
Real	11	16

Table 1. Confusion matrix for a domain expert distinguishing Real vs. Synthetic images.

Expert Turing Test: In this experiment, a domain expert (physician) was presented with a total of 54 samples in equal numbers of real and synthetic images in random order. Real labels of images are hidden. We ask to choose whether the given image looks

Tab. 1 shows domain expert's performance in distinguishing real from synthetic images. Out of 27 synthetic images, 12 were correctly identified, while 15 were mistaken as real. Among 27 real images, 11 were correctly



(a) Segmentation - Dice (b) Segmentation - Jaccard & AJI (c) Segmentation - HD (95) (d) Classification

Figure 1. **Comparison of segmentation and classification metrics on the PanNuke [4] dataset across augmentation scaling factors.** The addition of PathDIff-generated synthetic data consistently increases downstream classification and segmentation performance. 1x uses one *synthetic* augmentation set equal to the *real* train split size; 1.5x adds another *synthetic* set equal to 1.5 times the *real* train split size and so on. After 2.5x, the performance metrics plateau.

Algorithm 1 Unified Conditional Sampling

Require: ω: guidance strength
Define $\mathbf{c} \in \{(\varnothing_\mathbf{m}, \mathbf{c_t}), (\mathbf{c_m}, \varnothing_\mathbf{t}), (\mathbf{c_m}, \mathbf{c_t})\}$
$\{1, \ldots, T\}$: timesteps with decreasing noise schedule $\alpha = \{\alpha_t\}_{t=1}^{T}$
Initialize: $\mathbf{z_T} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T$ **to** 1 **do**
  ▷ Form classifier-free guided score at timestep $\mathbf{t}$
  $\tilde{\epsilon}_\theta(\mathbf{z_t}, \mathbf{t}, \mathbf{c}) = (\mathbf{1} + \omega)\epsilon_\theta(\mathbf{z_t}, \mathbf{t}, \mathbf{c}) - \omega\epsilon_\theta(\mathbf{z_t}, \mathbf{t})$
  ▷ Denoise step to obtain intermediate sample $\tilde{\mathbf{z}}_\mathbf{t}$
  $\tilde{\mathbf{z}}_\mathbf{t} = \frac{\mathbf{z_t} - \sqrt{\mathbf{1}-\bar{\alpha}_\mathbf{t}}\,\tilde{\epsilon}_\theta(\mathbf{z_t}, \mathbf{t}, \mathbf{c})}{\sqrt{\bar{\alpha}_\mathbf{t}}}$
  if $t > 1$ **then**
    Sample $\mathbf{z_{t-1}} \sim \mathcal{N}(\mu_\theta(\mathbf{z_t}, \tilde{\mathbf{z}}_\mathbf{t}, \mathbf{t}), \mathbf{\Sigma}_\theta(\mathbf{z_t}, \mathbf{t}))$
  else
    $\mathbf{z_0} = \tilde{\mathbf{z}}_\mathbf{t}$
  end if
end for
return $\mathbf{z_0}$

labeled, with 16 falsely classified as synthetic. The overall accuracy was approximately 42.6. This indicates the user found it somewhat challenging to differentiate real from synthetic images. *real* or *synthetic*.

4. Sampling Algorithm

We use classifier-free guidance to sample from conditional and unconditional diffusion models to update the final score. Algorithm 1 gives an overview of the sampling. We either randomly pair the conditions from non-overlapping M2I and T2I datasets or generate silver standard masks for T2I dataset (or can generate caption/relevant text condition for M2I dataset).

5. Considerations for p_{split}

When training jointly on two datasets—Text-to-Image and Mask-to-Image—p_{split} controls the proportion of data sampled from each of them. We evaluate performance with three values of p_{split}: 0.2, 0.5, and 0.8. Results using only text are shown in Table 2, mask-only conditioning in Tab. 3, and both text and mask conditioning in Tab. 4.

In these experiments, $p_{split} = 0.5$ strikes a balance, explaining why we chose this value in the main paper. While it seems logical to assign a larger probability to the larger dataset to cover more of its samples, we found that $p_{split} = 0.5$ works well in practice, ensuring samples from both datasets are included at least once per epoch.

p_{split}	PathCap: Train			PathCap: Test		
	FID ↓	KID ↓	PLIP ↑	FID ↓	KID ↓	PLIP ↑
$p_{split} = 0.2$	16.33	0.0624	24.43	15.74	0.0603	24.50
$p_{split} = 0.5$	18.52	0.0619	24.18	19.60	0.0644	24.05
$p_{split} = 0.8$	19.58	0.0649	24.34	18.87	0.0626	24.27

Table 2. **Considerations for p_{split}.** CLIP-FID [7, 13], KID [1], and PLIP [8] similarity scores for different p_{split} values on PathCap [15], with text condition c_t used for sampling. PLIP [8] similarity scores on the real PathCap train and test splits are **26.34** and **26.56**, respectively, provided as a reference for comparison.

p_{split}	PanNuke: Train		PanNuke: Test	
	FID ↓	KID ↓	FID ↓	KID ↓
$p_{split} = 0.2$	7.36	0.0525	7.88	0.0559
$p_{split} = 0.5$	6.94	0.0389	7.28	0.0415
$p_{split} = 0.8$	8.57	0.0584	8.97	0.0707

Table 3. **Considerations for p_{split}.** CLIP-FID [7, 13], KID [1] for different p_{split} values on PanNuke [4] dataset. Only mask condition c_m was used for sampling.

6. In-Domain FID Results

For a more faithful assessment of pathology image quality, we compute an in-domain FID using the CONCH [10] encoder rather than relying solely on CLIP or Inception-based features, which were trained on general natural images and may not capture the nuances of histopathology. CONCH [10] is a foundation model trained on large pathology image-text pairs.

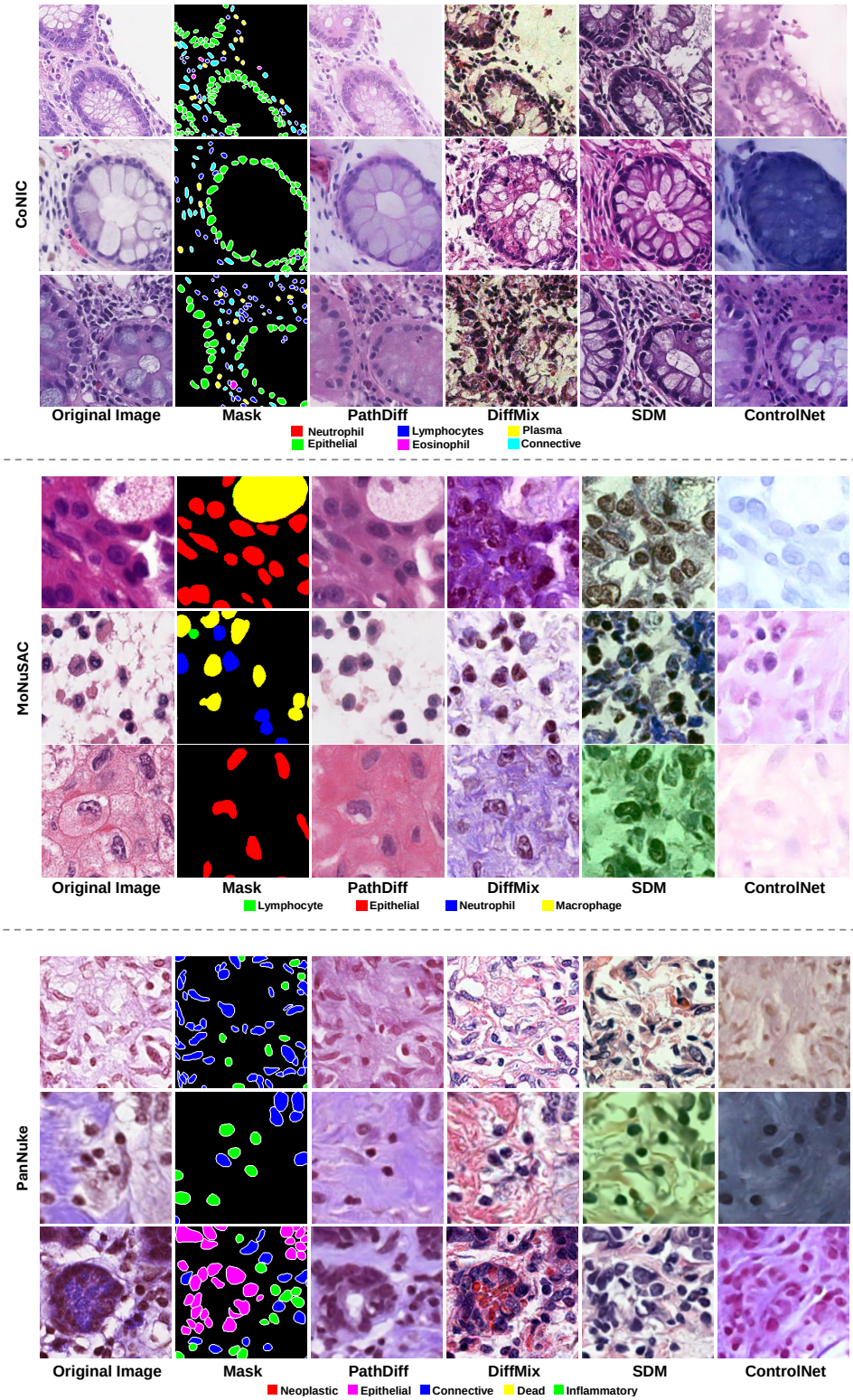


Figure 2. **Qualitative comparison** of synthetic images generated by PathDiff, DiffMix [12], SDM [17], and ControlNet [19] on the CoNIC [5], MoNuSAC [16], and PanNuke [4] datasets.

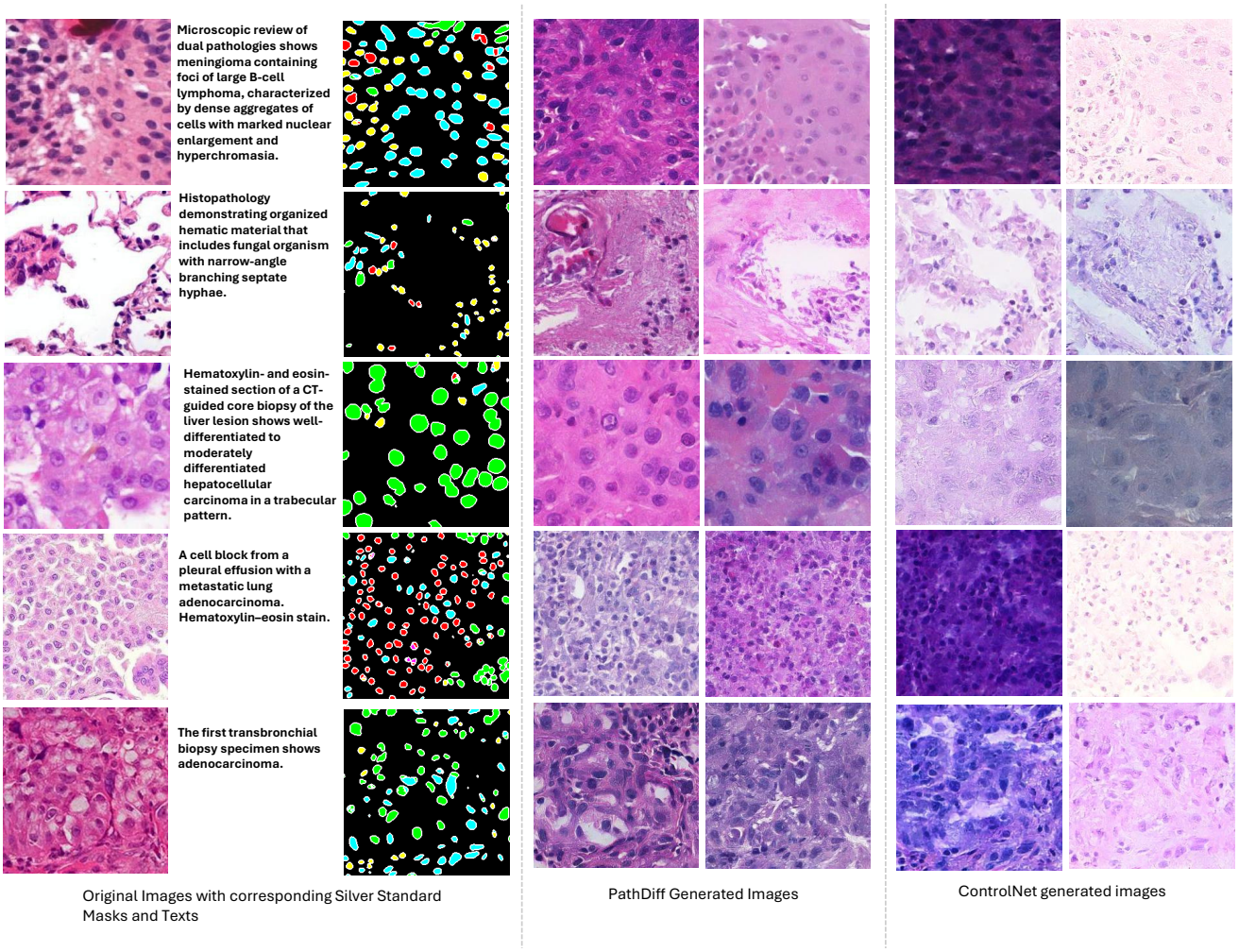


Figure 3. **Qualitative comparison of Pathdiff and ControlNet generated images on paired Silver standard masks and texts.**

p_{split}	PanNuke				PathCap					
	Train		Test		Train			Test		
	FID ↓	KID ↓	FID ↓	KID ↓	FID ↓	KID ↓	PLIP ↑	FID ↓	KID ↓	PLIP ↑
$p_{split} = 0.2$	10.25	0.0672	11.99	0.0800	15.21	0.0846	23.01	16.07	0.0956	22.81
$p_{split} = 0.5$	11.03	0.0718	12.32	0.0952	14.39	0.0884	23.01	14.26	0.1059	22.80
$p_{split} = 0.8$	9.723	0.0729	10.37	0.0862	12.78	0.0955	22.97	12.53	0.1107	22.70

Table 4. **Consideration for p_{split}.** CLIP-FID [7, 13], KID [7], and PLIP [8] similarity scores for different p_{split} values for PanNuke [4] and PathCap [15]. We used both text c_t and mask c_m for sampling.

7. Mask Depth Ablation

We test the effect of using two types of conditioning mask, first cell type mask and other is instance mask. We generate mask edges from instance mask using image processing technics. Using both masks generates better quality images as seen in Tab. 7, subsequently we use the mask depth of 6. We simply increase the channel size of the Mask embedder M and concatenate two masks as input.

8. Choice of Pretrained Checkpoints

We evaluated different pretrained checkpoints choices in three module components: VAE, Text-Encoder,and U-Net.

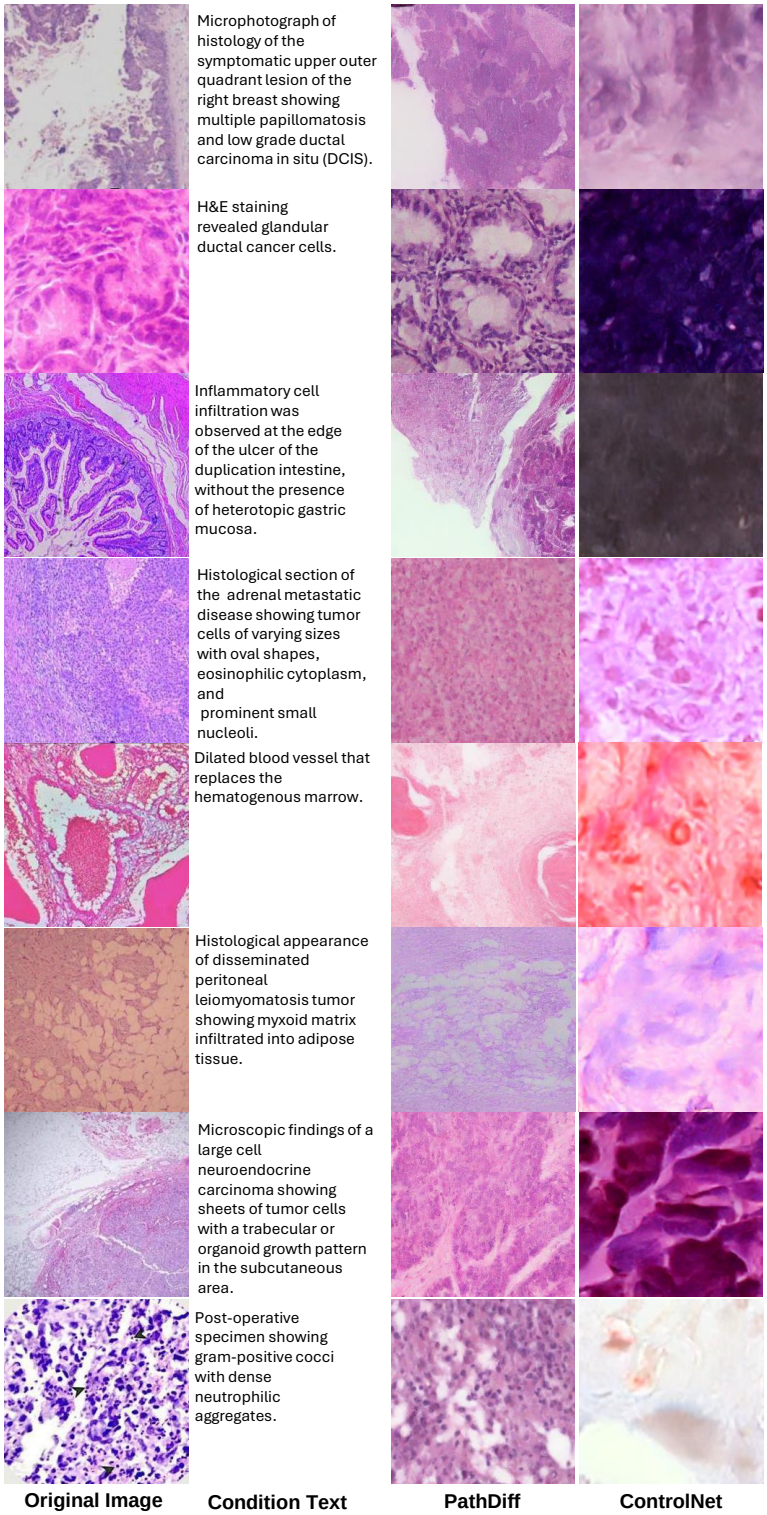


Figure 4. **Qualitative comparison** of synthetic images generated by PathDiff and ControlNet [19] on the PathCap [15] dataset.

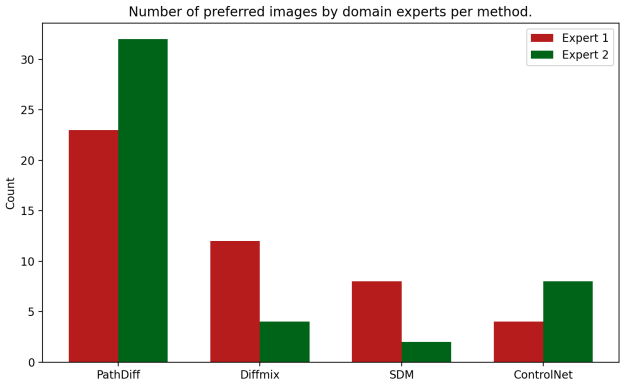


Figure 5. Both domain experts significantly preferred PathDiff generated images over other methods.

Method	PanNuke	CoNIC	MoNuSAC
Diffmix	119.35	257.92	290.27
SDM	177.30	143.48	166.65
ControlNet	121.57	174.74	277.96
PathDiff	**53.74**	**91.21**	**121.61**

Table 5. **Comparison of CONCH-FID** across training splits for PanNuke [4], CoNIC [5], and MoNuSAC [16]. PathDiff is trained jointly with T2I dataset: PathCap [15]. ControlNet [19] uses SD [14] backbone trained on the PathCap dataset.

Method	PanNuke	CoNIC	MoNuSAC
ControlNet	347.77	343.39	331.61
PathDiff	**153.77**	**156.21**	**141.88**

Table 6. **Comparison of CONCH-FID** on training splits for T2I dataset: PathCap. PathDiff is jointly trained with three M2I datasets: PanNuke [4], CoNIC [5], and MoNuSAC [16].

# mask_depth	PanNuke Test			
	IP ↑	IR ↑	CONCH FID ↓	KID ↓
3	0.79	0.47	102.37	0.0644
6	0.72	0.77	7.21	0.0415

Table 7. **Mask depth ablation** on PanNuke test split

Finetuning VAE The reconstruction performance of VAEs [14, 18] plays a crucial role in the fidelity of generated images. Losses introduced during the compression and decompression stages in VAEs compound with the denoising process losses in subsequent stages, directly impacting the quality of the generated images.

Initially, we used the VQ-VAE from [18], which was trained on the TCGA-BRCA [3] dataset containing whole-slide images (WSIs) exclusively from breast tissues. While this VAE outperforms the one from [14], which was trained on natural images, its applicability is limited as it lacks representation of diverse tissue types. We fine-tuned the VAE on the datasets used in this work, including PanNuke [4], PathCap [15], CoNIC [5], and MoNuSAC [16].

As demonstrated in Tab. 8, fine-tuning the VAE on these datasets results in improvements across all reconstruction and generation metrics. However, these improvements, while consistent, are relatively modest.

VAE Trained on	Metrics			
	LPIPS ↓	SSIM ↑	MSE ↓	FID ↓
TCGA-BRCA [3]	0.0462	0.7962	0.0084	6.94
Datasets: D	0.0429	0.8212	0.0070	6.31

Table 8. **Effect of fine-tuning VAE on datasets D:** PanNuke [4], PathCap [15], MoNuSAC [16], and CoNIC [5].

Text Encoder To evaluate text–image alignment on the PathCap training set, we compared the similarity between each image and its corresponding report using two embedding methods. The CLIP-based similarity score was 21.56, while the PLIP-based score reached 26.30; clearly demonstrating that PLIP embeddings achieve stronger alignment between images and text. Therefore we used PLIP text encoder checkpoint in our experiments.

Denoising U-net We fine-tune the U-Net on the PathCap Text-to-Image dataset starting from the TCGA-BRCA checkpoint provided by [6]. The results are summarized in Tab. 9. We see improved CLIP-FID and KID scores for Mask-to-Image generation on PanNuke, as well as a higher PLIP similarity score for Text-to-Image generation on PathCap.

U-Net Checkpoint	CLIP FID ↓	PLIP Score ↑	KID ↓
TCGA-BRCA	14.44	21.79	0.1284
PathCap	7.21	24.05	0.0410

Table 9. **Comparison of U-Net checkpoints** on CLIP FID, PLIP score, and KID

9. Significance Test on Downstream Task:

We validate that PathDiff's higher downstream scores aren't due to chance by running a paired permutation test. We randomly swap method labels within each test example and compare mean F1/Dice across 1000 trials. The resulting p-values are < 0.05, confirming PathDiff's gains are statistically significant as seen in Fig. 6.

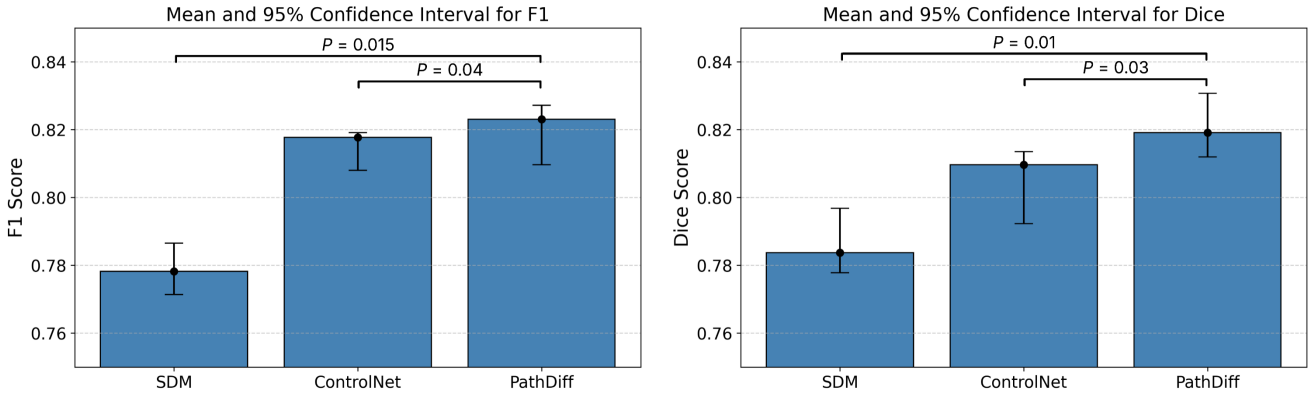


Figure 6. **Pairwise significance test**, $p < 0.05$ indicates that PathDiff augmentation set helps improve downstream classification and segmentation tasks statistically significantly as compared to other methods.

10. Generation Performance on Hard Pathology Cases

To assess how well PathDiff handles challenging, clinically significant images, we split our test set into "pathological" (reports mentioning "carcinoma") and "non-pathological" cases (reports describing benign findings). Table 10 compares FID, KID, and PLIP scores for each group. Although overall image fidelity remains similar, lower PLIP scores for pathological cases suggest that images with malignant features are marginally more difficult to synthesize than benign ones.

Table 10. Performance on pathological vs. non-pathological cases

Case Type	FID ↓	KID ↓	PLIP ↑
Pathological	19.97	0.04	23.15
Non-pathological	20.10	0.06	24.13

11. Details on training previous works:

Diffmix and SDM are trained on M2I datasets only. We use their official repositories to refer to their code. For both DiffMix and SDM we use same training settings for all M2I datasets that of PanNuke [4] in [12]. For ControlNet we pre-trained SD[14] model on T2I data first and then used only M2I data for finetuning, as recommended in the official controlnet tutorial.

12. Computational Costs:

Since PathDiff only trains U-Net encoder and shallow mask embedder, training costs remain modest, even for joint training. PathDiff trains 694 M parameters. Training time for the largest dataset combination (PathCap [15] + PanNuke [4]) is 30 Hours on 4 NVIDIA A6000 GPUs. Sampling 6,300(train split of PanNuke) images takes 3.5-4.5 hours.

13. Survey Tool:

We used an interactive web-based tool to conduct a domain expert's survey. Clear instructions were given to evaluate the images. Fig. 7 and Fig. 8 show the web interface used for the domain expert image preference experiment and the Turing test respectively.

References

[1] Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018. 2

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[3] JN Cancer Genome Atlas Research Network et al. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120, 2013. 6

[4] Jevgenij Gamper, Navid Alemi Koohbanani, Ksenija Benet, Ali Khuram, and Nasir Rajpoot. Pannuke: An open pan-cancer histology dataset for nuclei instance segmentation and classification. In *Digital Pathology*, pages 11–19, Cham, 2019. Springer International Publishing. 1, 2, 3, 4, 6, 7

[5] Simon Graham, Quoc Dang Vu, Mostafa Jahanifar, Martin Weigert, Uwe Schmidt, Wenhua Zhang, Jun Zhang, Sen Yang, Jinxi Xiang, Xiyue Wang, et al. Conic challenge: Pushing the frontiers of nuclear detection, segmentation, classification and counting. *Medical image analysis*, 92: 103047, 2024. 1, 3, 6

[6] Alexandros Graikos, Srikar Yellapragada, Minh-Quan Le, Saarthak Kapse, Prateek Prasanna, Joel Saltz, and Dimitris Samaras. Learned representation-guided diffusion models for large-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8532–8542, 2024. 6

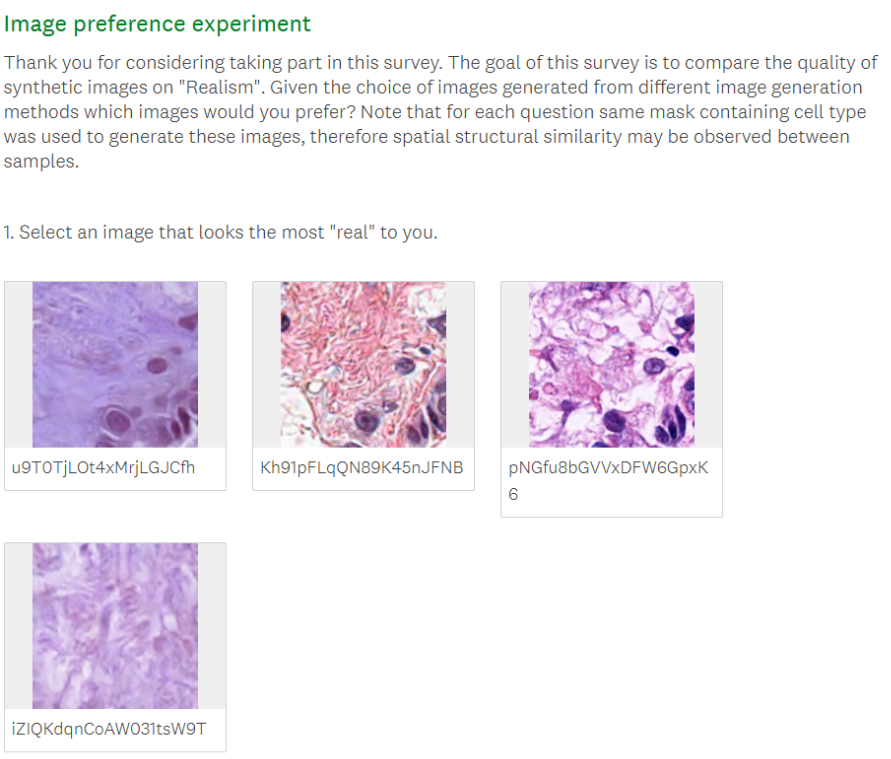


Figure 7. Interactive web interface used for domain expert image preference experiment.

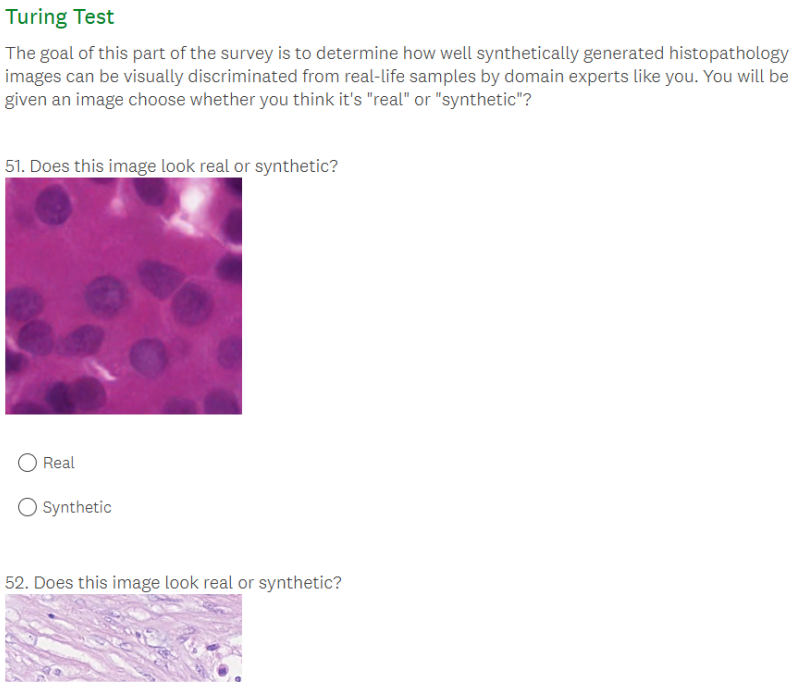


Figure 8. Interactive web interface used for domain expert Turing test.

[7] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 1, 2, 4

[8] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature Medicine*, pages 1–10, 2023. 2, 4

[9] Fabian Hörst, Moritz Rempe, Lukas Heine, Constantin Seibold, Julius Keyl, Giulia Baldini, Selma Ugurel, Jens Siveke, Barbara Grünwald, Jan Egger, and Jens Kleesiek. Cellvit: Vision transformers for precise cell segmentation and classification. *Medical Image Analysis*, 94:103143, 2024. 1

[10] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30:863–874, 2024. 2

[11] Seonghui Min, Hyun-Jic Oh, and Won-Ki Jeong. Co-synthesis of histopathology nuclei image-label pairs using a context-conditioned joint diffusion model. In *Computer Vision – ECCV 2024*, pages 146–162, Cham, 2025. Springer Nature Switzerland. 1

[12] Hyun-Jic Oh and Won-Ki Jeong. Diffmix: Diffusion model-based data synthesis for nuclei segmentation and classification in imbalanced pathology image datasets. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023: 26th International Conference, Vancouver, BC, Canada, October 8–12, 2023, Proceedings, Part III*, page 337–345, Berlin, Heidelberg, 2023. Springer-Verlag. 1, 3, 7

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. 2, 4

[14] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 6, 7

[15] Yuxuan Sun, Chenglu Zhu, Sunyi Zheng, Kai Zhang, Lin Sun, Zhongyi Shui, Yunlong Zhang, Honglin Li, and Lin Yang. Pathasst: A generative foundation ai assistant towards artificial general intelligence of pathology. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38: 5034–5042, 2024. 2, 4, 5, 6, 7

[16] Ruchika Verma, Neeraj Kumar, Abhijeet Patil, Nikhil Cherian Kurian, Swapnil Rane, and Amit Sethi. Multi-organ nuclei segmentation and classification challenge 2020. *IEEE transactions on medical imaging*, 39 (1380-1391):8, 2020. 1, 3, 6

[17] Weilun Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Dong Chen, Lu Yuan, and Houqiang Li. Semantic image synthesis via diffusion models, 2022. 1, 3

[18] Srikar Yellapragada, Alexandros Graikos, Prateek Prasanna, Tahsin Kurc, Joel Saltz, and Dimitris Samaras. Pathldm: Text conditioned latent diffusion model for histopathology. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5182–5191, 2024. 6

[19] Lvmin Zhang, Maneesh K. Wu, Weiyang Zeng, Yuxin Zhang, Hussain Salman, and Vladlen Koltun. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 1, 3, 5, 6