D-VST: Diffusion Transformer for Pathology-Correct Tone-Controllable Cross-Dye Virtual Staining of Whole Slide Images

Shurong Yang[1]* **Dong Wei**[2]* **Yihuang Hu**[1] **Qiong Peng**[1]
Hong Liu[1,2]† **Yawen Huang**[2] **Xian Wu**[2]‡ **Yefeng Zheng**[2,3]
Liansheng Wang[1]‡
[1]Xiamen University, Xiamen, China [2]Tencent Jarvis Lab, Shenzhen, China
[3]Westlake University, Hangzhou, China
{xuancheng, huyihuang, qpeng, liuhong}@stu.xmu.edu.cn
{donwei, yawenhuang, kevinxwu, yefengzheng}@tencent.com
lswang@xmu.edu.cn

Abstract

Diffusion-based virtual staining methods of histopathology images have demonstrated outstanding potential for stain normalization and cross-dye staining (e.g., hematoxylin-eosin to immunohistochemistry). However, achieving pathology-correct cross-dye virtual staining with versatile tone controls poses significant challenges due to the difficulty of decoupling the given pathology and tone conditions. This issue would cause non-pathologic regions to be mistakenly stained like pathologic ones, and vice versa, which we term "pathology leakage." To address this issue, we propose diffusion virtual staining Transformer (D-VST), a new framework with versatile tone control for cross-dye virtual staining. Specifically, we introduce a pathology encoder in conjunction with a tone encoder, combined with a two-stage curriculum learning scheme that decouples pathology and tone conditions, to enable tone control while eliminating pathology leakage. Further, to extend our method for billion-pixel whole slide image (WSI) staining, we introduce a novel frequency-aware adaptive patch sampling strategy for high-quality yet efficient inference of ultra-high resolution images in a zero-shot manner. Integrating these two innovative components facilitates a pathology-correct, tone-controllable, cross-dye WSI virtual staining process. Extensive experiments on three virtual staining tasks that involve translating between four different dyes demonstrate the superiority of our approach in generating high-quality and pathologically accurate images compared to existing methods based on generative adversarial networks and diffusion models. Our code and trained models are available at https://github.com/yangshurong/D-VST.

1 Introduction

Histological stainings are used to colorize tissue specimens, making the near-transparent tissue sections visible for pathological observations in clinical diagnostics and research [41]. Different types of dyes manifest different colors in stained tissue and provide complementary information; for example, hematoxylin-eosin (HE) can delineate the cellular structures, whereas immunohistochemistry (IHC)

*Co-first authors with equal contribution.
†Work done during an internship at Tencent Jarvis Lab.
‡Corresponding authors.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

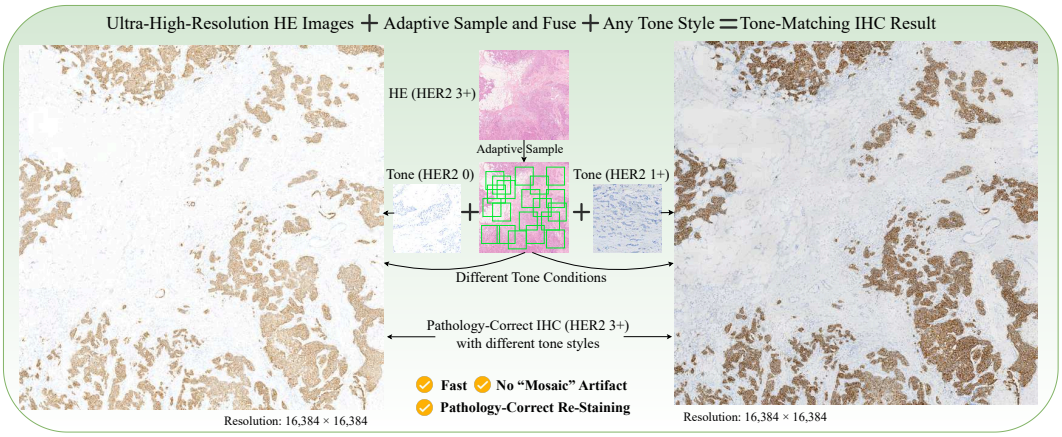

Figure 1: D-VST facilitates efficient (cf. runtime comparison in Table A10), high-quality, tone-controllable, and pathology-correct virtual staining of ultra-high-resolution histopathology images, featuring **adaptive patch sampling** to reduce inference overhead while eliminating mosaic artifacts (cf. Figure 6 and Figure A12); **versatile tone control** by different tone-conditioning images; and **correct pathological status** despite the status of the conditioning images (cf. quantitative and qualitative analysis in Table 3, Figure A10, and Figure A11). HER2 scores: 0: no cancerous lesion, and 1+, 2+, and 3+: increasing severity of cancerous lesions.

renders protein-specific expression to assist in tumor diagnosis and cancer prognosis [2]. However, current chemical protocols allow only one staining per tissue section; additional tissue sections are required for multiple stainings. This adds to the consumption of often limited tissue samples in clinics. In addition, the staining process is time- and chemical-consuming. Therefore, multiple stainings are resource-/labor-intensive and costly [2, 15, 36, 74, 77].

Virtual staining [2] provides a potential solution to multiple stainings—a cost-effective alternative to the conventional chemical process. It digitally "translates" chemically stained histopathology images using computational methods. Researchers leveraged generative adversarial networks (GANs) [37, 100] for virtual staining. Despite notable progress, GANs may encounter significant training challenges, such as mode collapse [79]. Recently, diffusion models have demonstrated superior quality to GANs in controllable image generation [9, 60, 67, 90, 91] and started to be applied to virtual staining of histopathology images. These applications can be divided into two groups: same-dye stain normalization and cross-dye staining. The former addresses the appearance variations in images stained with the same dye [38, 39, 70], likely originating from variations in institute, chemical material, or manual operation [12, 78, 80]. However, the generalization of these methods to the latter—image translation between two different dyes—remains to be investigated.

Cross-dye virtual staining translates histopathology images stained with one dye (the source domain) to new images that look like chemically stained with another (the target domain), e.g., HE to IHC, ideally without structure distortion or pathology status alteration. However, existing methods [20, 31, 33, 40, 53, 54, 86] cannot control the staining tones in the target domain,[4] leading to unpredictable randomness and significant variations in the tones of the virtually stained images. A potential solution is to condition the staining process [67, 95] with desirable tones. However, it is challenging to describe the tones with text. Also, providing the tone condition with an image is more complex than giving structure conditions with Canny edges. Specifically, the tone-conditioning images often contain mixed tone and pathology information. For example, in Figure 2(a), using a cancerous IHC image to condition the virtual staining of a cancer-free HE image may result in erroneous staining that falsely implies the presence of cancer pathology. We term this issue *pathology leakage*. To realize effective tone conditioning without pathology leaks, decoupling the tone and pathology information is crucial (Figure 2(b)).

This work presents diffusion virtual staining Transformer (D-VST), a diffusion model with a Transformer backbone for cross-dye virtual staining of histopathology images (Figure 1). D-VST controls

[4] In this work, we refer to the primary color of a type of dye (e.g., the generic pink color of HE staining) as **hue**, and the color variations of that dye as **tones** (e.g., dark to bright pink).

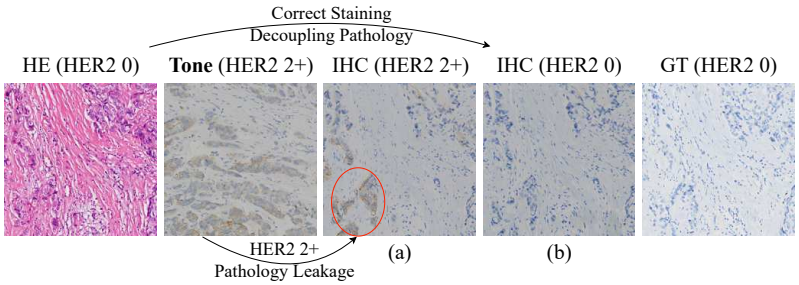

Figure 2: Illustration of **pathology leakage**. (a) Without special treatment, conditioning the virtual staining of a cancer-free HE image with a HER2 2+ IHC image may cause a leak of the cancerous status from the tone-conditioning image to the virtually stained image, leading to pathology-faulty staining. (b) Pathology-correct staining can be achieved by decoupling the tone and pathology conditions. HER2 scores: 0: no cancerous lesion, and 1+, 2+, and 3+: increasing severity of cancerous lesions. GT: ground truth.

the staining tone in the target domain without pathology leaks by adopting separate pathology and tone conditions. Concretely, it relies on the source-domain image to be re-stained for pathological structure conditioning and a target-domain image for tone conditioning. To our knowledge, D-VST represents the first endeavor to realize tone-controllable cross-dye virtual staining of histopathology images with diffusion models. In addition, we design a two-stage curriculum learning scheme to effectively decouple the model's learning of the pathology and tone conditions in a progressive manner. The first stage, pathology extraction, focuses on learning to extract pathology structure information from source-domain images without injecting tone control signals. Then, the second stage, tone injection, adds tone control while diminishing pathology information from the target-domain tone-conditioning image. This involves applying random dropout and Gaussian blur to the tone-conditioning image.

In addition, the virtual staining of large histopathology images like whole slide images (WSIs) requires processing ultra-high-resolution data. However, due to hardware constraints, directly denoising an entire WSI is challenging for diffusion models. Current methods typically divide a WSI into patches, process them individually, and then stitch them together [1, 39, 45, 59, 64]. As the patches are processed independently, this workaround often leads to discrepancies in color, brightness, and contrast between the stitched patches—resulting in the "mosaic" artifact [75]. The mosaic artifact may harm or even invalidate the clinical usability of the virtually stained WSI. To address a similar artifact of content discontinuity in a text-to-panorama application, MultiDiffusion [3] proposed denoising highly overlapping image patches yielded by a sliding-window process separately, followed by fusing the denoising directions by averaging the denoised patches within the overlapped regions. However, unlike natural panoramas, WSIs present significant variations in information density across an image's regions. As a result, the uniform sliding windows in MultiDiffusion may be sub-optimal for WSI virtual staining.

In this work, we present an efficient and high-quality zero-shot inference strategy for virtual staining of WSIs using diffusion models trained under a prevalent resolution, e.g., 512×512 pixels. Our observation indicates that the mosaic artifact is more pronounced in low-frequency regions of the virtually stained images. Leveraging this insight, we devise a frequency-aware adaptive patch sampling strategy to improve the generation quality of low-frequency regions while controlling computational overhead in high-frequency areas. This strategy enables efficient and rapid virtual staining of billion-pixel WSIs without notable mosaic artifacts, significantly enhancing the capability of our proposed D-VST framework.

Our contributions are summarized as follows:

- We propose D-VST, a novel Diffusion Transformer (DiT) [60] based model for histopathology image virtual staining. So far as we know, D-VST is the first diffusion model that realizes tone control for cross-dye virtual staining.
- To address the unwanted pathology leakage issue accompanying the tone control, we design an effective, two-step curriculum learning scheme with separate conditioning branches for pathology and tone.

- We propose an adaptive frequency-aware patch sampling strategy for efficient and high-quality zero-shot staining of billion-pixel WSIs.
- Last but not least, we conduct extensive experiments on three virtual staining tasks involving four dyes to evaluate our D-VST against up-to-date approaches. We also assess a downstream task and perform ablation studies on our method.

2 Related work

GAN-based virtual staining. Conventional methods predominantly employed GANs [7, 94, 37, 100] for virtual staining of histopathology images. A large amount of work [5, 8, 21, 47, 48, 49, 50, 59, 61, 81, 82, 85, 88] facilitated the transfer of HE to IHC images. [1, 34, 41, 65, 66] showcased generating HE images from formalin fixation and paraffin embedding (FFPE) ones. Moreover, [11, 13, 69] implemented GAN-based image style transfer for stain normalization of histopathology images, effectively mitigating color variations. However, GANs are known to be subject to the mode collapse issue [79] and challenging to train. The emerging diffusion models have recently demonstrated superior training stability, generation controllability, image quality, and versatility to GANs.

Diffusion-based virtual staining. Recent advancements in diffusion models [9, 60, 67, 90, 91] have showcased impressive controllable generation capabilities in image synthesis tasks. Various studies [38, 39, 70] employed diffusion models for *stain normalization of histopathology images*. StainDiff [70] proposed self-supervision to facilitate one-to-one color style transfer. StainFuser [39] leveraged the ControlNet [95] to implement fast neural style transfer. [40, 53, 54, 86] examined the potential of diffusion models for *cross-dye histopathology image virtual staining*. [31, 33] implemented cross-dye virtual staining without relying on pathological category labels. VIMs [20] introduced text-controlled protein markers to facilitate virtual staining across multiple pathological categories. However, these methods cannot control the tones for virtual staining, resulting in unpredictable appearance variations among the re-stained images. In contrast, our method controls the staining tone with a target-domain tone-conditioning image.

WSI generation. Virtual staining of WSIs, characterized by ultra-high resolution, presents significant challenges to diffusion models. [26, 30, 35, 52, 71, 72, 84, 87, 89, 96] introduced additional global control signals for direct high-resolution image generation by diffusion models. Yet, this approach does not apply to WSIs due to computational limitations. Alternatively, the sliding window strategy [3, 17, 19, 24, 28, 46, 75] offers a viable means. However, this strategy is impeded by substantial inference times due to the highly redundant sliding windows with small sliding steps, or subject to a performance drop in image quality with large sliding steps. [22, 99] proposed selecting patches at varying timesteps to mitigate long inference duration, yet this mechanism may introduce instability to the generated outcome. To facilitate efficient and high-quality WSI virtual staining, we propose a novel adaptive patch sampling strategy based on image frequency variations.

3 Method

The framework of our method is shown in Figure 3 (left). Primarily, D-VST tailors and extends the Diffusion Transformer (DiT) [60] in PixArt-α [9] as its denoising Transformer. A pathology encoder encodes the source image into pathological structure embeddings, which are injected into the denoising Transformer as the structure condition after concatenating with the noise latent. Meanwhile, a tone encoder (which we use the pretrained Vision Transformer (ViT) [18] in CLIP [63]) encodes an auxiliary, tone-conditioning image from the target domain into tone embeddings. The tone embeddings are also injected into the denoising Transformer via multi-head tone attention. The denoising Transformer integrates the structure and tone conditions and denoises toward a re-stained target histopathology image (latent).

The pipeline of D-VST proposes innovative designs for both training and inference schemes. To prevent pathology leakage from the tone-conditioning image, we decompose the intricate virtual staining task into a curriculum learning [4] streamline for progressive training [68]. To eliminate the mosaic artifact at a low computational cost while staining high-resolution images, we propose a frequency-aware adaptive patch sampling strategy for efficient inference.

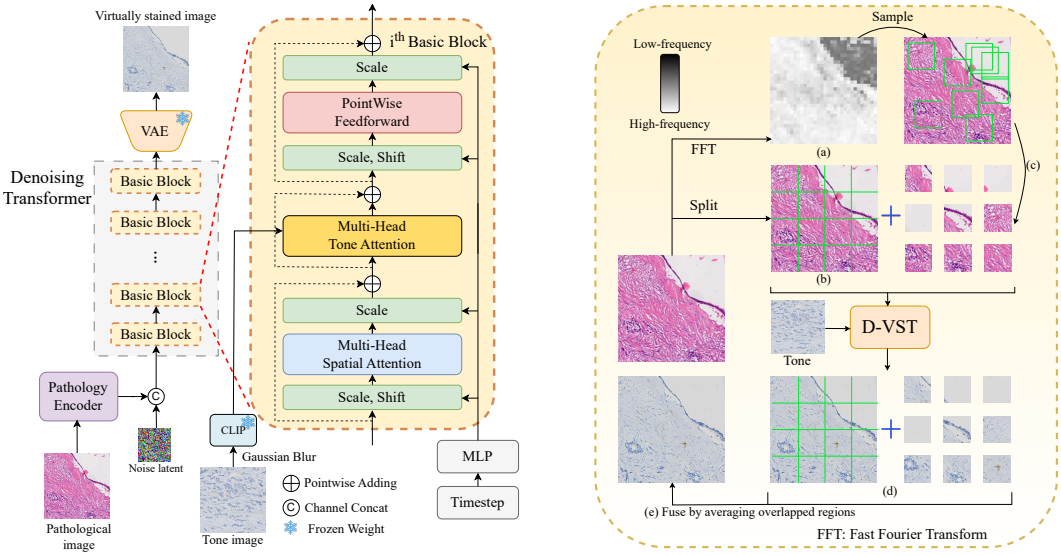

Figure 3: Left: **framework of the proposed D-VST**. Right: **frequency-aware adaptive patch sampling.** (a) FFT-based local frequency computing. (b) Covering the entire image with tiled, non-overlapping patches. (c) Additional patches sampled according to local frequency. We show only nine patches here to illustrate that patches are more likely to be sampled from low- than high-frequency regions. In fact, all sampled patches cover seven times the area of the input image. (d) Individually denoising each patch. (e) Fusing by averaging overlapped regions.

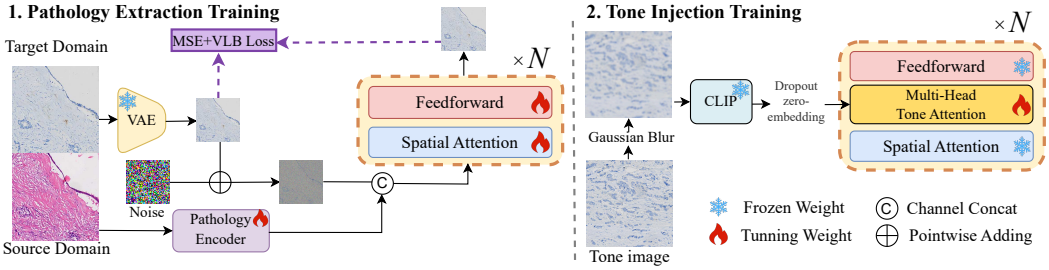

Figure 4: The **two-stage curriculum learning scheme** decouples pathology and tone conditions. Initially, only the pathology condition is input to the model, emphasizing accurate pathology extraction while ignoring the significant hue difference between the source and target domains. Next, we add tone condition via cross-attention, with random dropout and Gaussian blur to diminish pathology leakage from the tone-conditioning image.

3.1 Curriculum learning scheme decoupling pathology and tone conditions

As introduced in Figure 2, training with mingled pathology and tone controls often causes pathological status to leak through tone conditioning, leading to incorrect staining outcomes. This occurs because the model confuses the purposes of the conditioning images and mistakes the tone-conditioning image for the source of the pathology status. Thus, the key to effectively preventing pathology leakage is to decouple pathology and tone conditions by making the model learn both control signals precisely. For this purpose, we design a novel two-stage curriculum learning [4] scheme. In the first stage, pathology conditioning is first learned alone. Then, tone conditioning is introduced in the second stage for further joint training.

Training stage 1: pathology extraction. Figure 4 (left) illustrates the first training stage. In this stage, we only feed the pathology condition into the model via channel-wise concatenation with the noise latent. Meanwhile, the denoising Transformer includes only the spatial attention and feedforward modules. The training goal is to denoise the corrupted target-domain image (obtained by adding Gaussian noise as in typical diffusion processes) in the latent space of a pretrained variational

autoencoder (VAE)[5] [43], guided by the pathology structure embedding extracted from the *paired* pathology-conditioning source-domain image. We follow [60] to train the model with the hybrid mean squared error (MSE) and variational lower bound (VLB) losses [55], where the former learns to predict the sampled noise and the latter learns variances of the reverse diffusion process (cf. Appendix for more details). Thus, the model learns to effectively extract and utilize the complex pathology information in the pathology-conditioning image in this stage, ignoring the significant hue difference between the source and target domains.

The pathology encoder uses a lightweight convolutional network with 4×4 kernels, 2×2 strides, and four layers of 16, 32, 64, and 128 channels. It encodes the pathology-conditioning image into an embedding of the same shape as the noise latent. The embedding is concatenated with the noise latent along the channel dimension and input to the denoising Transformer. Compared to the ControlNet [95] architecture used in StainFuser [39], our lightweight pathology encoder is equally effective in capturing detailed pathological information while substantially reducing model complexity and computational cost.

Training stage 2: tone injection. The second stage introduces tone conditioning into the model (Figure 4 (right)). A random patch from the same WSI but *different from (thus not paired with)* the target-domain image is used for tone conditioning. It provides a precise tone style but not necessarily pathology information of the target image (different patches of a WSI may present distinct pathological statuses). To emphasize tone features while minimizing pathological structural information, we first apply a Gaussian blur to the tone-conditioning image. Then, we utilize the pretrained ViT [18] in OpenAI-CLIP [63] to encode the blurred image into a tone embedding. The OpenAI-CLIP ViT was trained on massive data of versatile *colors*, making it a proper *tone* encoder for *optical* pathology images (cf. Appendix for a comparison to the PathCLIP [76]). Next, in the middle of the frozen feedforward and spatial attention modules, we insert a multi-head tone attention module into each unit block of the denoising Transformer trained in the first stage. Lastly, using cross-attention, we inject the tone embedding into the denoising Transformer via the inserted tone attention modules. To further reduce pathology leakage, we apply a random dropout to the tone embedding by replacing it with a zero embedding. The dropout rate is set to 20% according to preliminary trials.[6] The same training losses as in the first stage are used. Ablation studies confirm the efficacy of the tone encoder's components (Table 4).

3.2 Frequency-aware adaptive patch sampling

From the "No-overlap" column of Figure 6, we observe that the mosaic artifact is more pronounced in low-frequency regions of the virtually stained images. This occurs because, unlike areas with complex textures that contain abundant clues guiding the virtual staining process, low-frequency regions require more overlapping patches for consistent denoising results. To improve the generation quality of the low-frequency regions while simultaneously controlling computational overhead, we propose a frequency-aware adaptive patch sampling strategy. As the premise, we divide the input image into a grid of $n \times n$ squares. We set $n = 32$ through a grid search (cf. Appendix). Then, for each square I, we compute the natural logarithm of the magnitudes of its fast Fourier transform (FFT; Figure 3 (right)-(a)):

$$L = \log\big[\,\mathrm{abs}\big(\,\mathrm{FFT}(I)\big)\big]. \tag{1}$$

Next, we calculate the pixel-wise mean of L, denoted by l, as the frequency statistic for the square. Finally, we convert l to sampling probability by:

$$p_i = \frac{(l^{\max} + l^{\min} - l_i)^{\alpha}}{\sum_{i=1}^{n^2}(l^{\max} + l^{\min} - l_i)^{\alpha}}, \tag{2}$$

where $l^{\max}$ and $l^{\min}$ are the maximum and minimum l values of all n^2 squares, $(l^{\max} + l^{\min} - l_i)$ makes low-frequency squares more likely to be sampled, and $\alpha \in \mathbb{Z}^+$ is a hyperparameter controlling the difference in sampling probabilities between low- and high-frequency squares.

[5] In our preliminary experiments, when applying the VAE of Stable Diffusion (SD) 1.5 [67] to pathology image reconstruction, we obtained an average peak signal-to-noise ratio of ∼28 dB, comparable to the performance on natural images (∼25 dB). Therefore, we use the SD 1.5 VAE in this work.

[6] Further optimization via rigorous ablation study may lead to better performance; we leave it for future work.

To ensure the entire input image is stained, we first fully cover the image with tiled, non-overlapping patches (Figure 3 (right)-(b); note the patches are larger than the squares for FFT). Next, we sample additional patches according to the sampling probabilities in Equation (2) (Figure 3 (right)-(c)): a square I is first selected according to p_i, and a random pixel within I is subsequently chosen as the center of an additional patch. This way, more patches are adaptively sampled where necessary, whereas fewer patches are sampled from high-frequency regions to maintain reasonable computational costs. In this work, we define an integer β as the ratio of the total area of tiled and sampled patches to the area of the stained image. The larger β is, the more patches are sampled (the special case of no-overlap sliding windows has $\beta = 1$). The patches are processed separately in every denoising step, and the denoised patches are fused in overlapping areas by averaging (Figure 3 (right)-(d) and (e)) [3]. Experiments show that compared with MultiDiffusion [3], our strategy improves the virtual staining quality with only 12.5% computational overhead ($\beta = 8$ versus 64; cf. Table 2).

4 Experiments

Datasets and evaluation metrics. We evaluate D-VST on three datasets to comprehensively validate its performance on the virtual staining of various dyes. RegH2I [61] comprises 2,592 pairs of registered images for HE to IHC staining (HE2IHC), [34] includes 5,098 pairs of aligned images for FFPE to HE (FFPE2HE) staining, and HEMIT [5] contains 5,292 matched image pairs for HE to multiplex immunohistochemistry (mIHC) staining (HE2mIHC). These datasets include two organs/cancer types: breast cancer (HE2IHC and FFPE2HE) and colon cancer (HE2mIHC). We use all datasets' official train/test/validation splits. Unless otherwise specified, we report performance on cropped images of 1024×1024 pixels as in [5, 34, 61], and use a random patch from the same WSI but *not overlapping* with the target image for tone conditioning. Note that the tone-conditioning and target images may present different pathological statuses, and the no-overlap requirement ensures no structural leak.

Following previous works [5, 34, 61], we employ the structural similarity index measure (SSIM) and peak signal-to-noise ratio (PSNR) for quality assessment of the virtually stained images. However, for paired histopathology images stained with different dyes, perfect pixel-to-pixel matching is practically impossible even after registration (see Appendix for more explanation). Therefore, for an appropriate evaluation, we additionally employ three metrics that are more perceptually relevant than the conventional SSIM and PSNR: deep image structure and texture similarity (DISTS) [16], Fréchet inception distance (FID), and kernel inception distance (KID) [6].

Implementation details. All experiments are conducted in Python 3.10.0 with PyTorch 2.0.0 [58] on a GPU with 80 GB of memory. We follow [9] to use DiT-XL/2 [60] as the base network architecture for our denoising Transformer, and the pretrained parameters from [9]. We employ the AdamW [44] optimizer with a learning rate of 10^{-5} and a batch size 32. We train for 30,000 steps for the pathology extraction stage, and an additional 10,000 steps for the tone injection stage. The diffusion time T is set to 1000. Our model trains and infers at the resolution of 512×512 pixels. For virtual staining of larger images, we use the model to denoise 512×512 patches sampled according to the proposed frequency-aware adaptive patch sampling strategy (cf. Section 3.2), and fuse the patch-wise outcomes by averaging overlapped regions [3]. Unless otherwise specified, we set $\alpha = 1$ and $\beta = 8$ for the adaptive sampling. Our code and trained models are available at https://github.com/yangshurong/D-VST.

Comparison with state-of-the-art (SOTA). We compare our D-VST with classical GAN-based image translation methods: CycleGAN [100], pix2pix [37] and pix2pixHD [83]; medical image diffusion model: SynDiff [56]; and SOTA GAN/diffusion models specialized in histopathology image virtual staining: [34] and [61]/StainFuser [39]. The comparisons with [34] and [61] are exclusively on the FFPE2HE [34] and HE2IHC [61] datasets, respectively, since the two methods were designed for the specific tasks. As shown in Table 1, D-VST achieves the best performance for all metrics on the HE2IHC and HE2mIHC datasets, indicating that the histopathology images virtually stained by D-VST are superior both perceptually and structurally. On the FFPE2HE dataset, D-VST yields slightly inferior PSNR and SSIM to the best numbers, yet is still competitive. We conjecture this is because the micro-level correspondence (pixel- and structure-wise) between the paired images in this dataset is not as good as the other two. Notwithstanding, D-VST again achieves the best performance

Table 1: Evaluation of various methods on three different datasets and virtual staining tasks.

Method	DISTS↓	FID↓	KID↓	PSNR↑	SSIM↑
HE2IHC [61]					
Pix2pix [37]	0.192	47.77	0.0237	18.04±3.693	0.401±0.126
Pix2pixHD [83]	0.191	41.77	0.0123	18.06±3.733	0.386±0.128
CycleGAN [100]	0.212	40.91	0.0062	17.01±3.524	0.365±0.119
[61]	0.174	33.92	0.0058	18.02±3.706	0.385±0.125
SynDiff [56]	0.348	225.3	0.2282	18.09±4.049	0.404±0.114
StainFuser [39]	0.255	104.5	0.0791	17.30±3.949	0.401±0.148
D-VST	**0.154**	**33.16**	**0.0055**	**18.11±3.874**	**0.407±0.136**
HE2mIHC [5]					
Pix2pix[37]	0.133	29.95	0.0058	27.26±3.903	0.855±0.063
Pix2pixHD [83]	0.170	28.92	0.0086	27.65±3.916	0.816±0.062
CycleGAN [100]	0.300	83.16	0.0365	20.15±1.663	0.520±0.049
SynDiff [56]	0.318	316.1	0.4057	20.05±1.423	0.709±0.038
StainFuser [39]	0.289	99.20	0.0690	20.27±2.087	0.309±0.040
D-VST	**0.106**	**20.36**	**0.0016**	**28.01±4.123**	**0.861±0.045**
FFPE2HE [34]					
Pix2pix [37]	0.109	17.88	0.0008	18.34±2.009	0.536±0.113
Pix2pixHD [83]	0.091	15.26	0.0008	19.08±2.046	0.586±0.106
CycleGAN [100]	0.171	35.36	0.0087	14.47±2.045	0.363±0.152
[34]	0.126	31.56	0.0101	**19.86±2.082**	**0.644±0.098**
SynDiff [56]	0.228	69.44	0.0329	12.15±1.787	0.336±0.146
StainFuser [39]	0.200	64.98	0.0301	12.46±1.851	0.263±0.157
D-VST	**0.090**	**14.26**	**0.0005**	17.98±2.165	0.538±0.117

Table 2: Evaluation of sampling strategies for zero-shot staining of images larger than the training resolution of diffusion models. β is the ratio of the total patch area to the area of the virtually stained image. Given a fixed patch size, the larger β is, the more patches are sampled, thus the higher computational cost.

Sample strategy	β	DISTS↓	FID↓	KID↓	PSNR↑	SSIM↑
HE2IHC [61]						
No-overlap	1	0.1594	37.748	0.0092	18.01±3.868	0.406±0.136
SpotDiffusion [22]	1	0.2256	93.349	0.0664	14.08±2.815	0.297±0.117
MultiDiffusion [3]	64	0.1566	34.661	0.0061	18.09±3.877	0.406±0.136
D-VST	8	**0.1548**	**33.162**	**0.0055**	**18.11±3.874**	**0.407±0.136**
HE2mIHC [5]						
No-overlap	1	0.1063	21.866	0.0021	27.72±4.534	0.852±0.048
SpotDiffusion [22]	1	0.4285	111.35	0.0937	10.87±3.717	0.373±0.059
MultiDiffusion [3]	64	0.1069	20.365	0.0016	27.96±4.260	0.858 ±0.045
D-VST	8	**0.1062**	**20.361**	**0.0016**	**28.01±4.123**	**0.861±0.045**
FFPE2HE [34]						
No-overlap	1	0.0929	16.403	0.0012	17.60±2.183	0.517±0.120
SpotDiffusion [22]	1	0.1483	48.281	0.0256	15.41±1.408	0.467±0.109
MultiDiffusion [3]	64	0.0910	14.960	0.0007	17.87±2.165	0.531±0.118
D-VST	8	**0.0900**	**14.263**	**0.0005**	**17.98±2.166**	**0.538±0.117**

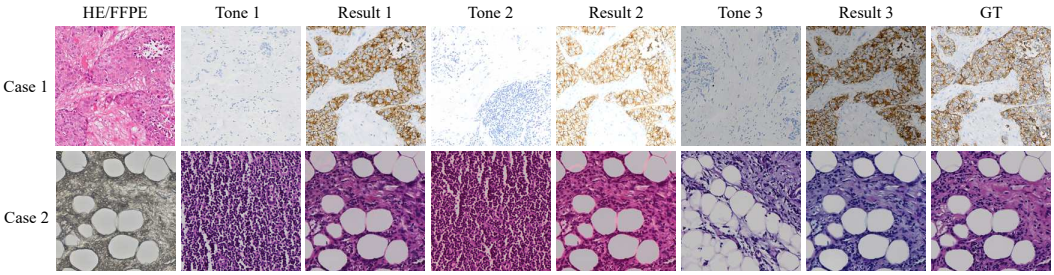

Figure 5: Qualitative results for **tone control**. Case 1 is from the HE2IHC [61] dataset, whereas Case 2 is from the FFPE2HE [34] dataset. Despite substantial discrepancies in pathological status between the pathology- and tone-conditioning images, the pathological status is correctly transferred from the pathology-conditioning images to the virtually stained ones. For example, in Case 1, the HER2 score of the HE image is 3+, while the scores of Tone 1, 2, and 3 are all 0. GT: ground truth.

for the three perception-oriented metrics (DISTS, FID, and KID). In general, D-VST demonstrates strong capabilities in virtually staining high-quality histopathology images of versatile dyes.

Tone control and downstream task. As the HE2mIHC [5] dataset has undergone color normalization and thus cannot provide varying tone conditions, we conduct qualitative tone control experiments on the other two datasets. As shown in Figure 5, when conditioned on histopathology images of various tones of another dye, the virtually stained images exhibit varying tones matching the tone-conditioning images while maintaining the same pathological status as the source pathology-conditioning images. For example, Case 1 illustrates that even when there are substantial discrepancies in pathological status between the pathology- and tone-conditioning images, the pathological status is still correctly transferred from the pathology-conditioning image to the virtually stained ones. These observations indicate that our D-VST can effectively prevent pathology leakage for tone-conditioned cross-dye virtual staining. We provide more visualizations and comparisons with other methods in Appendix.

To further *quantitatively* validate our method's effectiveness in pathology leakage prevention while using image-based tone conditioning, we perform a downstream classification task on HE2IHC [61]. Concretely, we further split the 600 official test pairs into a sub-train and sub-test set of 480 and 120 pairs, respectively. Then, we train a ResNet50 [29] classifier on the IHC images of the sub-train set, with labels corresponding to the four HER2 scores (HER2 0: no cancerous lesions, and 1+, 2+, and 3+: increasing severity of cancerous lesions, with higher scores indicating more pronounced lesions and more advanced disease stages). Next, for each HE image in the sub-test set, we randomly select an IHC image in the sub-test set that is not paired with the specific HE image as the tone condition for virtual staining. Lastly, we apply the trained classifier to the virtually stained IHC

Table 3: Evaluation of downstream classification task on the HE2IHC [61] dataset.

Method	ACC↑	F1↑	Precision↑	Recall↑
Pix2pix [37]	0.8750	0.8755	0.8916	0.8779
Pix2pixHD [83]	0.8500	0.8508	0.8655	0.8384
CycleGAN [100]	0.5583	0.5609	0.5600	0.5512
[61]	0.8917	0.8925	0.8930	0.8847
SynDiff [56]	0.2167	0.0884	0.3025	0.2589
StainFuser [39]	0.7250	0.7324	0.7331	0.7230
D-VST	**0.9417**	**0.9430**	**0.9470**	**0.9388**
Real IHC images	0.9500	0.9506	0.9530	0.9488

Table 4: Ablation study on the HE2IHC [61] dataset with both image generation and downstream task metrics.

Metric	w/o Dropout	w/o Gaussian	w/o Curriculum	D-VST
ACC↑	0.9250	0.7833	0.7667	**0.9417**
F1↑	0.9265	0.7710	0.7672	**0.9430**
Precision↑	0.9292	0.8425	0.7720	**0.9470**
Recall↑	0.9250	0.7585	0.7645	**0.9388**
DISTS↓	0.1589	0.2372	0.1729	**0.1548**
FID↓	35.815	55.148	38.957	**33.162**
KID↓	0.0070	0.0290	0.0110	**0.0055**
PSNR↑	17.451±3.629	15.852±3.440	17.587±3.574	**18.113±3.874**
SSIM↑	0.3904±0.142	**0.4361±0.145**	0.4012±0.137	0.4074±0.136

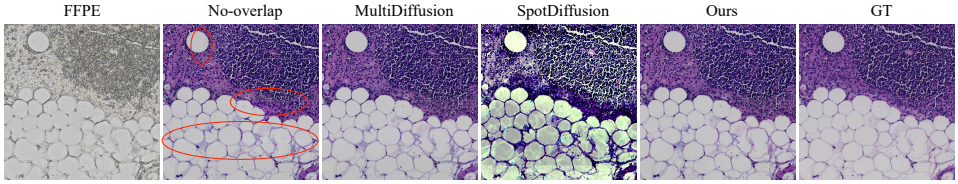

Figure 6: Comparison of sampling strategies for zero-shot virtual staining of large histopathology images. Red ellipses outline regions where the mosaic artifacts are prominent (best viewed zoomed in). GT: HE ground truth. More is provided in Appendix.

images. Intuitively, if the pathology leakage happens, the classification results would notably deviate from directly classifying the real IHC images in the sub-test set. Table 3 shows that our classifier achieves high performance on real IHC images, and more importantly, the performance on virtually stained IHC images by D-VST is also high and closely matches that on real ones. Meanwhile, D-VST obtains substantially better results than the compared methods. These results demonstrate that our method successfully decouples pathology and tone information for tone-conditioned cross-dye virtual staining, and its virtually stained images are of high quality for potential clinical use.

Sampling strategy for zero-shot virtual staining of large histopathology images. As described in the implementation details, our model trains and infers at the resolution of 512×512 pixels. For the virtual staining of larger images, we use the model to denoise 512×512 patch samples, followed by patch fusion. Here, we compare several sampling strategies in terms of performance and computational cost: no-overlap (sliding windows without overlap), SpotDiffusion [22] (sampling sliding windows that vary with timesteps), MultiDiffusion [3] (sliding windows with a high overlap ratio), and our proposed frequency-aware adaptive patch sampling. Since different sampling strategies are evaluated on the same network (our proposed), which takes about 1.081s and 6.4 GB memory to infer a patch of 512×512 pixels on our hardware, their relative computational costs can be compared by β values. We conduct quantitative evaluations on the HE2IHC, FFPE2HE, and HE2mIHC datasets, generating images of 1024×1024 pixels. As shown in Table 2, although no-overlap and SpotDiffusion incur the least computational cost, their performance is the worst. MultiDiffusion improves all evaluated metrics, though at 64 times the inference cost. In contrast, our strategy achieves the best performance for all metrics on the three datasets, while incurring only 1/8 of the inference computation of MultiDiffusion. For more insights, we additionally evaluate MultiDiffusion with the same $\beta = 8$ as ours on HE2IHC. Its FID degrades from 34.66 to 34.99, markedly inferior to our 33.16. These results demonstrate that our frequency-aware adaptive sampling strategy is not only highly efficient but also capable of boosting the quality of virtual staining.

Figure 6 shows example results by the compared methods for qualitative analysis, virtually stained at the resolution of 2048×2048 pixels. No-overlap exhibits noticeable mosaic artifacts, whereas SpotDiffusion presents anomalous tones—accounting for their unsatisfactory quantitative results. With only 1/8 of the computational cost of MultiDiffusion, our strategy produces images of equal visual quality to MultiDiffusion. We have also applied D-VST to the virtual staining of WSIs of 0.2–1.3 billion pixels (16,000×15,000 to 40,000×32,000 pixels). However, as far as we know, no suitable WSI dataset is currently available for reliable quantitative evaluation at scale. Therefore, we only show the qualitative results in Appendix for an observational study.

Ablation studies. In Section 3.1, we have proposed two-stage curriculum learning, Gaussian blur of tone-conditioning images, and random dropout of the tone condition to realize pathology and

tone decoupling and prevent pathology leakage. Here, we ablate one of them at a time (denoted by w/o Curriculum, w/o Gaussian, and w/o Dropout) to study their efficacy on the HE2IHC dataset using both image generation and downstream classification metrics. Table 4 shows that removing any of them leads to overall declines in all metrics (the only exception is the SSIM w/o Gaussian). Especially, removing curriculum learning results in the most significant performance drops in three of the four classification metrics. These results suggest that these components, especially the two-stage curriculum learning scheme, effectively boost the performance of histopathology image virtual staining. This is achieved by improving the perceptual quality via effective pathology and tone condition disentanglement, thus fulfilling our design.

The Appendix includes further experiments determining the values of the hyperparameters α and β in the frequency-aware adaptive patch sampling (cf. Section 3.2).

5 Conclusion

This work presented D-VST, a diffusion Transformer based framework for efficient, high-quality, and pathology-preserving cross-dye virtual staining of histopathology images with up to more than a billion pixels. Extensive experiments on three virtual staining tasks involving four types of dyes and a downstream cancer status classification task validated D-VST's promising performance. Facilitating efficient tone-controllable virtual staining, D-VST has the potential to make a broad impact on algorithm development and the clinical pipeline of histopathology image analysis.

Limitations and future work. Our D-VST facilitates efficient virtual staining of ultra-high-resolution histopathology images like WSIs by the proposed frequency-aware adaptive patch sampling strategy. However, its inference speed is still constrained by the multi-step denoising process inherent in diffusion models [32]. Inspired by [42, 51, 73], we plan to optimize the denoising scheduler and reduce inference steps by flow rectification and consistency models, and further reduce computation overhead and accelerate inference by model pruning and distillation retraining [93].

Obtaining paired cross-stain training data can be challenging and costly in real-world workflows, which may limit the scalability and applicability of D-VST. While such data can improve virtual staining performance with paired correspondence, unpaired data is substantially more scalable due to orders of magnitude larger amounts. In future work, we plan to explore benefiting from both the scalability of unpaired data and the quality of paired data via a combination of D-VST and approaches [92] like CycleGAN-Turbo [27, 57], which enable diffusion models to learn from unpaired data.

In this work, we have attempted to fine-tune the VAE alongside the Diffusion Transformer in our experiments but obtained mixed results (similar PSNR and SSIM with poorer DISTS, FID, and KID), likely due to the limited training data. In the future, we plan to explore whether substituting the VAE with a histopathology-image-pretrained counterpart would further enhance our framework's performance.

Lastly, future work will investigate D-VST's benefits for downstream segmentation tasks [97].

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62371409).

References

[1] Muhammad Zeeshan Asaf, Babar Rao, Muhammad Usman Akram, Sajid Gul Khawaja, Samavia Khan, Thu Minh Truong, Palveen Sekhon, Irfan J Khan, and Muhammad Shahmir Abbasi. Dual contrastive learning based image-to-image translation of unstained skin tissue into virtually stained H&E images. *Scientific Reports*, 14(1):2335, 2024.

[2] Bijie Bai, Xilin Yang, Yuzhu Li, Yijie Zhang, Nir Pillar, and Aydogan Ozcan. Deep learning-enabled virtual histological staining of biological samples. *Light: Science & Applications*, 12(1):57, 2023.

[3] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: fusing diffusion paths for controlled image generation. In *ICML*. JMLR.org, 2023.

[4] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *ICML*, pages 41–48, 2009.

[5] Chang Bian, Beth Philips, Tim Cootes, and Martin Fergie. HEMIT: H&E to multiplex-immunohistochemistry image translation with dual-branch pix2pix generator. *arXiv preprint arXiv:2403.18501*, 2024.

[6] Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *ICLR*, 2018.

[7] Jiezhang Cao, Langyuan Mo, Yifan Zhang, Kui Jia, Chunhua Shen, and Mingkui Tan. Multi-marginal Wasserstein GAN. *NeurIPS*, 32, 2019.

[8] Fuqiang Chen, Ranran Zhang, Boyun Zheng, Yiwen Sun, Jiahui He, and Wenjian Qin. Pathological semantics-preserving learning for H&E-to-IHC virtual staining. In *MICCAI*, pages 384–394. Springer, 2024.

[9] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

[10] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Bowen Chen, Andrew Zhang, Daniel Shao, Andrew H Song, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 2024.

[11] Hyungjoo Cho, Sungbin Lim, Gunho Choi, and Hyunseok Min. Neural stain-style transfer learning using GAN for histopathological images. *arXiv preprint arXiv:1710.08543*, 2017.

[12] Francesco Ciompi, Oscar Geessink, Babak Ehteshami Bejnordi, Gabriel Silva De Souza, Alexi Baidoshvili, Geert Litjens, Bram Van Ginneken, Iris Nagtegaal, and Jeroen Van Der Laak. The importance of stain normalization in colorectal tissue classification with convolutional networks. In *IEEE ISBI*, pages 160–163. IEEE, 2017.

[13] Cong Cong, Sidong Liu, Antonio Di Ieva, Maurice Pagnucco, Shlomo Berkovsky, and Yang Song. Colour adaptive generative networks for stain normalisation of histopathology images. *Medical Image Analysis*, 82:102580, 2022.

[14] Timothy Michael D'Alfonso, Yi-Fang Liu, Zhengming Chen, Ying-Bei Chen, Ashley Cimino-Mathews, and Sandra Jean Shin. SP3, a reliable alternative to herceptest in determining HER-2/neu status in breast cancer patients. *Journal of Clinical Pathology*, 66(5):409–414, 2013.

[15] Astrid De Cuyper, Marc Van Den Eynde, and Jean-Pascal Machiels. HER2 as a predictive biomarker and treatment target in colorectal cancer. *Clinical Colorectal Cancer*, 19(2):65–72, 2020.

[16] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE TPAMI*, 44(5):2567–2581, 2020.

[17] Zheng Ding, Mengqi Zhang, Jiajun Wu, and Zhuowen Tu. Patched denoising diffusion models for high-resolution image synthesis. In *ICLR*, 2023.

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[19] Ruoyi Du, Dongliang Chang, Timothy Hospedales, Yi-Zhe Song, and Zhanyu Ma. DemoFusion: Democratising high-resolution image generation with no $$$. In *CVPR*, pages 6159–6168, 2024.

[20] Shikha Dubey, Yosep Chong, Beatrice Knudsen, and Shireen Y Elhabian. VIMs: Virtual immunohistochemistry multiplex staining via text-to-stain diffusion trained on uniplex stains. *arXiv preprint arXiv:2407.19113*, 2024.

[21] Shikha Dubey, Tushar Kataria, Beatrice Knudsen, and Shireen Y Elhabian. Structural cycle GAN for virtual immunohistochemistry staining of gland markers in the colon. In *International Workshop on Machine Learning in Medical Imaging*, pages 447–456. Springer, 2023.

[22] Stanislav Frolov, Brian B Moser, and Andreas Dengel. SpotDiffusion: A fast approach for seamless panorama generation over time. *arXiv preprint arXiv:2407.15507*, 2024.

[23] Parmida Ghahremani, Joseph Marino, Ricardo Dodds, and Saad Nadeem. DeepLIIF: An online platform for quantification of clinical pathology slides. In *CVPR*, pages 21399–21405, 2022.

[24] 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 *CVPR*, pages 8532–8542, 2024.

[25] Erik Großkopf, Valay Bundele, Mehran Hossienzadeh, and Hendrik Lensch. Histdist: histopathological diffusion-based stain transfer. *arXiv preprint arXiv:2505.06793*, 2025.

[26] Lanqing Guo, Yingqing He, Haoxin Chen, Menghan Xia, Xiaodong Cun, Yufei Wang, Siyu Huang, Yong Zhang, Xintao Wang, Qifeng Chen, et al. Make a cheap scaling: A self-cascade diffusion model for higher-resolution adaptation. *arXiv preprint arXiv:2402.10491*, 2024.

[27] Yong Guo, Jian Chen, Jingdong Wang, Qi Chen, Jiezhang Cao, Zeshuai Deng, Yanwu Xu, and Mingkui Tan. Closed-loop matters: Dual regression networks for single image super-resolution. In *CVPR*, pages 5407–5416, 2020.

[28] Moayed Haji-Ali, Guha Balakrishnan, and Vicente Ordonez. Elasticdiffusion: Training-free arbitrary size image generation through global-local content separation. In *CVPR*, pages 6603–6612, 2024.

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[30] Yingqing He, Shaoshu Yang, Haoxin Chen, Xiaodong Cun, Menghan Xia, Yong Zhang, Xintao Wang, Ran He, Qifeng Chen, and Ying Shan. ScaleCrafter: Tuning-free higher-resolution visual generation with diffusion models. In *ICLR*, 2023.

[31] Yufang He, Zeyu Liu, Mingxin Qi, Shengwei Ding, Peng Zhang, Fan Song, Chenbin Ma, Huijie Wu, Ruxin Cai, Youdan Feng, et al. PST-Diff: Achieving high-consistency stain transfer by diffusion models with pathological and structural constraints. *IEEE TMI*, 2024.

[32] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020.

[33] Man M Ho, Shikha Dubey, Yosep Chong, Beatrice Knudsen, and Tolga Tasdizen. F2FLDM: Latent diffusion models with histopathology pre-trained embeddings for unpaired frozen section to FFPE translation. *arXiv preprint arXiv:2404.12650*, 2024.

[34] Yihuang Hu, Qiong Peng, Zhicheng Du, Guojun Zhang, Huisi Wu, Jingxin Liu, Hao Chen, and Liansheng Wang. Boosting FFPE-to-HE virtual staining with cell semantics from pretrained segmentation model. In *MICCAI*, pages 67–76. Springer, 2024.

[35] Linjiang Huang, Rongyao Fang, Aiping Zhang, Guanglu Song, Si Liu, Yu Liu, and Hongsheng Li. FouriScale: A frequency perspective on training-free high-resolution image synthesis. *arXiv preprint arXiv:2403.12963*, 2024.

[36] Nida Iqbal and Naveed Iqbal. Human epidermal growth factor receptor 2 (HER2) in cancers: overexpression and therapeutic implications. *Molecular Biology International*, 2014(1):852748, 2014.

[37] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 1125–1134, 2017.

[38] Jiheon Jeong, Ki Duk Kim, Yujin Nam, Cristina Eunbee Cho, Heounjeong Go, and Namkug Kim. Stain normalization using score-based diffusion model through stain separation and overlapped moving window patch strategies. *Computers in Biology and Medicine*, 152:106335, 2023.

[39] Robert Jewsbury, Ruoyu Wang, Abhir Bhalerao, Nasir Rajpoot, and Quoc Dang Vu. StainFuser: Controlling diffusion for faster neural style transfer in multi-gigapixel histology images. *arXiv preprint arXiv:2403.09302*, 2024.

[40] Tushar Kataria, Beatrice Knudsen, and Shireen Y Elhabian. StainDiffuser: Multitask dual diffusion model for virtual staining. *arXiv preprint arXiv:2403.11340*, 2024.

[41] Umair Khan, Sonja Koivukoski, Mira Valkonen, Leena Latonen, and Pekka Ruusuvuori. The effect of neural network architecture on virtual H&E staining: Systematic assessment of histological feasibility. *Patterns*, 4(5), 2023.

[42] Bo-Kyeong Kim, Hyoung-Kyu Song, Thibault Castells, and Shinkook Choi. BK-SDM: A lightweight, fast, and cheap version of stable diffusion. In *ECCV*, pages 381–399. Springer, 2025.

[43] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[44] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[45] Sonja Koivukoski, Umair Khan, Pekka Ruusuvuori, and Leena Latonen. Unstained tissue imaging and virtual hematoxylin and eosin staining of histologic whole slide images. *Laboratory Investigation*, 103(5):100070, 2023.

[46] Yuseung Lee, Kunho Kim, Hyunjin Kim, and Minhyuk Sung. SyncDiffusion: Coherent montage via synchronized joint diffusions. *NeurIPS*, 36:50648–50660, 2023.

[47] Fangda Li, Zhiqiang Hu, Wen Chen, and Avinash Kak. Adaptive supervised PatchNCE loss for learning HE-to-IHC stain translation with inconsistent groundtruth image pairs. In *MICCAI*, pages 632–641. Springer, 2023.

[48] Jiahan Li, Jiuyang Dong, Shenjin Huang, Xi Li, Junjun Jiang, Xiaopeng Fan, and Yongbing Zhang. Virtual immunohistochemistry staining for histological images assisted by weakly-supervised learning. In *CVPR*, pages 11259–11268, 2024.

[49] Shengjie Liu, Chuang Zhu, Feng Xu, Xinyu Jia, Zhongyue Shi, and Mulan Jin. BCI: Breast cancer immunohistochemical image generation through pyramid pix2pix. In *CVPR*, pages 1815–1824, 2022.

[50] Shuting Liu, Baochang Zhang, Yiqing Liu, Anjia Han, Huijuan Shi, Tian Guan, and Yonghong He. Unpaired stain transfer using pathology-consistent constrained generative adversarial networks. *IEEE TMI*, 40(8):1977–1989, 2021.

[51] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

[52] Xinyu Liu, Yingqing He, Lanqing Guo, Xiang Li, Bu Jin, Peng Li, Yan Li, Chi-Min Chan, Qifeng Chen, Wei Xue, et al. HiPrompt: Tuning-free higher-resolution generation with hierarchical MLLM prompts. *arXiv preprint arXiv:2409.02919*, 2024.

[53] 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 *ECCV*, pages 146–162. Springer, 2025.

[54] Puria Azadi Moghadam, Sanne Van Dalen, Karina C Martin, Jochen Lennerz, Stephen Yip, Hossein Farahani, and Ali Bashashati. A morphology focused diffusion probabilistic model for synthesis of histopathology images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2000–2009, 2023.

[55] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, pages 8162–8171. PMLR, 2021.

[56] Muzaffer Özbey, Onat Dalmaz, Salman UH Dar, Hasan A Bedel, Şaban Özturk, Alper Güngör, and Tolga Çukur. Unsupervised medical image translation with adversarial diffusion models. *IEEE TMI*, 2023.

[57] Gaurav Parmar, Taesung Park, Srinivasa Narasimhan, and Jun-Yan Zhu. One-step image translation with text-to-image models. *arXiv preprint arXiv:2403.12036*, 2024.

[58] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. *NeurIPS*, 32, 2019.

[59] Pushpak Pati, Sofia Karkampouna, Francesco Bonollo, Eva Compérat, Martina Radić, Martin Spahn, Adriano Martinelli, Martin Wartenberg, Marianna Kruithof-de Julio, and Marianna Rapsomaniki. Accelerating histopathology workflows with generative AI-based virtually multiplexed tumour profiling. *Nature Machine Intelligence*, pages 1–17, 2024.

[60] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4195–4205, 2023.

[61] Qiong Peng, Weiping Lin, Yihuang Hu, Ailisi Bao, Chenyu Lian, Weiwei Wei, Meng Yue, Jingxin Liu, Lequan Yu, and Liansheng Wang. Advancing H&E-to-IHC virtual staining with task-specific domain knowledge for HER2 scoring. In *MICCAI*, pages 3–13. Springer, 2024.

[62] Colin A Purdie, Lee B Jordan, Jean B McCullough, Sharon L Edwards, Joan Cunningham, Miriam Walsh, Andrew Grant, Norman Pratt, and Alastair M Thompson. HER2 assessment on core biopsy specimens using monoclonal antibody CB11 accurately determines HER2 status in breast carcinoma. *Histopathology*, 56(6):702–707, 2010.

[63] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021.

[64] Aman Rana, Gregory Yauney, Alarice Lowe, and Pratik Shah. Computational histological staining and destaining of prostate core biopsy RGB images with generative adversarial neural networks. In *International Conference on Machine Learning and Applications*, pages 828–834. IEEE, 2018.

[65] Yair Rivenson, Tairan Liu, Zhensong Wei, Yibo Zhang, Kevin de Haan, and Aydogan Ozcan. PhaseStain: the digital staining of label-free quantitative phase microscopy images using deep learning. *Light: Science & Applications*, 8(1):23, 2019.

[66] Yair Rivenson, Hongda Wang, Zhensong Wei, Kevin de Haan, Yibo Zhang, Yichen Wu, Harun Günaydın, Jonathan E Zuckerman, Thomas Chong, Anthony E Sisk, et al. Virtual histological staining of unlabelled tissue-autofluorescence images via deep learning. *Nature Biomedical Engineering*, 3(6):466–477, 2019.

[67] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022.

[68] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

[69] Pegah Salehi and Abdolah Chalechale. Pix2pix-based stain-to-stain translation: A solution for robust stain normalization in histopathology images analysis. In *International Conference on Machine Vision and Image Processing*, pages 1–7. IEEE, 2020.

[70] Yiqing Shen and Jing Ke. StainDiff: Transfer stain styles of histology images with denoising diffusion probabilistic models and self-ensemble. In *MICCAI*, pages 549–559. Springer, 2023.

[71] Shuwei Shi, Wenbo Li, Yuechen Zhang, Jingwen He, Biao Gong, and Yinqiang Zheng. ResMaster: Mastering high-resolution image generation via structural and fine-grained guidance. *arXiv preprint arXiv:2406.16476*, 2024.

[72] Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. FreeU: Free lunch in diffusion U-net. In *CVPR*, pages 4733–4743, 2024.

[73] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *ICML*, ICML'23. JMLR.org, 2023.

[74] Edward C Stack, Chichung Wang, Kristin A Roman, and Clifford C Hoyt. Multiplexed immunohistochemistry, imaging, and quantitation: a review, with an assessment of tyramide signal amplification, multispectral imaging and multiplex analysis. *Methods*, 70(1):46–58, 2014.

[75] Kexin Sun, Zhineng Chen, Gongwei Wang, Jun Liu, Xiongjun Ye, and Yu-Gang Jiang. Bidirectional feature fusion generative adversarial network for ultra-high resolution pathological image virtual re-staining. In *CVPR*, pages 3904–3913, 2023.

[76] 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. In *AAAI*, volume 38, pages 5034–5042, 2024.

[77] Sandra M Swain, Mythili Shastry, and Erika Hamilton. Targeting HER2-positive breast cancer: advances and future directions. *Nature Reviews Drug Discovery*, 22(2):101–126, 2023.

[78] David Tellez, Geert Litjens, Péter Bándi, Wouter Bulten, John-Melle Bokhorst, Francesco Ciompi, and Jeroen Van Der Laak. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical Image Analysis*, 58:101544, 2019.

[79] Hoang Thanh-Tung and Truyen Tran. Catastrophic forgetting and mode collapse in GANs. In *International Joint Conference on Neural Networks*, pages 1–10. IEEE, 2020.

[80] Quoc Dang Vu, Robert Jewsbury, Simon Graham, Mostafa Jahanifar, Shan E Ahmed Raza, Fayyaz Minhas, Abhir Bhalerao, and Nasir Rajpoot. Nuclear segmentation and classification: on color and compression generalization. In *International Workshop on Machine Learning in Medical Imaging*, pages 249–258. Springer, 2022.

[81] Qiuli Wang, Yongxu Liu, Li Ma, Xianqi Wang, Wei Chen, and Xiaohong Yao. Histology virtual staining with mask-guided adversarial transfer learning for tertiary lymphoid structure detection. *arXiv preprint arXiv:2408.13978*, 2024.

[82] Song Wang, Zhong Zhang, Huan Yan, Ming Xu, and Guanghui Wang. Mix-domain contrastive learning for unpaired H&E-to-IHC stain translation. *arXiv preprint arXiv:2406.11799*, 2024.

[83] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional GANs. In *CVPR*, pages 8798–8807, 2018.

[84] Xiaojuan Wang, Janne Kontkanen, Brian Curless, Steven M Seitz, Ira Kemelmacher-Shlizerman, Ben Mildenhall, Pratul Srinivasan, Dor Verbin, and Aleksander Holynski. Generative powers of ten. In *CVPR*, pages 7173–7182, 2024.

[85] Linda Wei, Shengyi Hua, Shaoting Zhang, and Xiaofan Zhang. Derestainer: H&E to IHC pathological image translation via decoupled staining channels. In *MICCAI Workshop on Deep Generative Models*, pages 1–10. Springer, 2024.

[86] Dominik Winter, Nicolas Triltsch, Marco Rosati, Anatoliy Shumilov, Ziya Kokaragac, Yuri Popov, Thomas Padel, Laura Sebastian Monasor, Ross Hill, Markus Schick, et al. Mask-guided cross-image attention for zero-shot in-silico histopathologic image generation with a diffusion model. *arXiv preprint arXiv:2407.11664*, 2024.

[87] Haoning Wu, Shaocheng Shen, Qiang Hu, Xiaoyun Zhang, Ya Zhang, and Yanfeng Wang. MegaFusion: Extend diffusion models towards higher-resolution image generation without further tuning. *arXiv preprint arXiv:2408.11001*, 2024.

[88] Zhaoyang Xu, Xingru Huang, Carlos Fernández Moro, Béla Bozóky, and Qianni Zhang. GAN-based virtual re-staining: a promising solution for whole slide image analysis. *arXiv preprint arXiv:1901.04059*, 2019.

[89] Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and CUI Bin. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal LLMs. In *ICML*, 2024.

[90] Shurong Yang, Huadong Li, Juhao Wu, Minhao Jing, Linze Li, Renhe Ji, Jiajun Liang, and Haoqiang Fan. MegActor: Harness the power of raw video for vivid portrait animation. *arXiv preprint arXiv:2405.20851*, 2024.

[91] Shurong Yang, Huadong Li, Juhao Wu, Minhao Jing, Linze Li, Renhe Ji, Jiajun Liang, Haoqiang Fan, and Jin Wang. Megactor-σ: Unlocking flexible mixed-modal control in portrait animation with diffusion transformer. *arXiv preprint arXiv:2408.14975*, 2024.

[92] Yifan Yang, Shuhai Zhang, Zixiong Huang, Yubing Zhang, and Mingkui Tan. Cross-ray neural radiance fields for novel-view synthesis from unconstrained image collections. In *ICCV*, pages 15901–15911, 2023.

[93] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *CVPR*, pages 6613–6623, 2024.

[94] Cheng Yu, Wenmin Wang, and Roberto Bugiolacchi. Improving generative adversarial network inversion via fine-tuning gan encoders. *Applied Soft Computing*, 166:112201, 2024.

[95] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023.

[96] Shen Zhang, Zhaowei Chen, Zhenyu Zhao, Zhenyuan Chen, Yao Tang, Yuhao Chen, Wengang Cao, and Jiajun Liang. HiDiffusion: Unlocking high-resolution creativity and efficiency in low-resolution trained diffusion models. *arXiv preprint arXiv:2311.17528*, 2023.

[97] Yichi Zhang, Zhenrong Shen, and Rushi Jiao. Segment anything model for medical image segmentation: Current applications and future directions. *Computers in Biology and Medicine*, 171:108238, 2024.

[98] Yijie Zhang, Luzhe Huang, Nir Pillar, Yuzhu Li, Hanlong Chen, and Aydogan Ozcan. Pixel super-resolved virtual staining of label-free tissue using diffusion models. *Nature Communications*, 16(1):5016, 2025.

[99] Qingping Zheng, Yuanfan Guo, Jiankang Deng, Jianhua Han, Ying Li, Songcen Xu, and Hang Xu. Any-size-diffusion: Toward efficient text-driven synthesis for any-size HD images. In *AAAI*, volume 38, pages 7571–7578, 2024.

[100] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2223–2232, 2017.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction are consistent with the paper's contributions and scope. They outline the methodologies and key findings presented in the study.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: Yes, we discuss the limitations of our work in the Appendix.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper does not include theoretical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide all the information for reproducibility in the Experiments section (e.g., compute resources and experimental settings) and the Appendix (e.g., some hyperparameter settings).

   Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We will release the code upon paper acceptance, with sufficient instructions to faithfully reproduce the main experimental results.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We describe all the necessary details in Section 4, Experiments and the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report the standard deviations for sample-wise metrics (PSNR and SSIM).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide this information, e.g., type and number of GPU, time of inference in Section 4.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in this paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: No direct societal impact of the work expected.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: We anticipate no direct malicious use for an approach to virtual re-staining of histopathology images.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We have ensured that all assets used in our research are properly credited to their original creators.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: As stated in the Abstract and Section 4, we will release our code and trained models, plus detailed instructions on how to run the code to reproduce our results.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This study does not involve human subjects, and therefore no IRB approval is required.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A1 Training losses

We briefly describe the losses for training our framework. In denoising diffusion probabilistic models (DDPMs) [32], a forward process gradually applies noise to real data x_0: $q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I})$, where t is the time step and constants $\bar{\alpha}_t$ are hyper-parameters. Applying the reparameterization trick, x_t can be sampled by $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t$, where $\epsilon_t \sim \mathcal{N}(0, \mathbf{I})$. Inversely, a model learns the reverse process to gradually restore the noise-corrupted real data by $p_\theta(x_{t-1}|x_t) = \mathcal{N}(\mu_\theta(x_t), \Sigma_\theta(x_t))$, using neural networks to predict the statistics of p_θ. The model is trained with the variational lower bound (VLB) of the log-likelihood of x_0, which can be written as (omitting a training-irrelevant term): $\mathcal{L}_{\mathrm{VLB}} = -\log p_\theta(x_0|x_1) + \sum_t \mathcal{D}_{\mathrm{KL}}\big(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)\big)$, where $\mathcal{D}_{\mathrm{KL}}$ is the Kullback-Leibler (KL) divergence loss. By parameterizing μ_θ as a noise prediction network ϵ_θ, the model can be trained with the mean-squared error loss between the predicted noise $\epsilon_\theta(x_t)$ and the ground-truth sampled Gaussian noise ϵ_t: $\mathcal{L}_{\mathrm{MSE}} = \|\epsilon_\theta(x_t) - \epsilon_t\|_2^2$. Meanwhile, to learn the covariance Σ_θ, the full $\mathcal{L}_{\mathrm{VLB}}$ needs to be optimized. We follow [55] to train ϵ_θ with $\mathcal{L}_{\mathrm{MSE}}$, and Σ_θ with $\mathcal{L}_{\mathrm{VLB}}$. Since the training losses are not a focus of this paper, we refer interested readers to [55] for more details.

A2 Justification for non-perfect paired data

In virtual staining tasks, paired images are typically obtained from consecutive tissue sections and algorithmically registered. Although the alignment does not perfectly match all pixels, most are closely aligned and thus valid for structural correspondence learning. Before ours, many methods successfully trained their virtual staining models on the datasets used in this work [5, 34, 61].

In the main text, we conjectured that D-VST's lower PSNR/SSIM versus [34] on FFPE2HE "is because the micro-level correspondence (pixel- and structure-wise) between the paired images in this dataset is not as good as the other two." Although this misalignment is inherent in all datasets used in this work due to the consecutive slicing and chemical staining process, we visually find it more serious in the FFPE2HE dataset. To quantify the structural (mis)alignment between paired images, we resort to the following procedures. We apply the Canny edge detector and compute the Hausdorff distance, intersection-over-union (IoU), and Dice similarity between edge maps of the source and target images. Intuitively, lower Hausdorff distance and higher IoU and Dice metrics indicate better structural alignment. As shown in Table A5, FFPE2HE [34] shows consistently worse alignment metrics, supporting our conjecture. In particular, the Hausdorff distance measures the maximum deviation between two point sets, highlighting the worst-case alignment errors. Thus, the substantially larger Hausdorff distances indicate more extreme misalignments.

Table A5: Quantification of structural (mis)alignment between paired images in the HE2IHC and FFPE2HE datasets using Hausdorff distance, intersection-over-union (IoU), and Dice similarity.

Datasets	Hausdorff↓	IoU↑	Dice↑
HE2IHC	30.92±12.05	0.150±0.026	0.260±0.041
FFPE2HE	36.63±33.13	0.138±0.011	0.242±0.016

A3 Additional experiments

Choice of tone encoder. For the tone encoder, we experimented with three image encoders: one pretrained in OpenAI-CLIP [63] on massive data of broad spectrums; one pretrained in PathCLIP [76] on 207K high-quality pathology image–caption pairs; and one pretrained in UNI [10] on over 200 million pathology HE and IHC images. Table A6 presents their performance on the HE2IHC dataset [61]. UNI and OpenAI-CLIP demonstrate comparable performance, and clearly outperform PathCLIP in perception-oriented metrics while remaining comparable in PSNR and SSIM. We conjecture that the difference in performance may be partly attributed to the function of the tone encoder. On the one hand, while PathCLIP may be better prepared for downstream tasks on pathology images (e.g., classification), the tone encoder focuses more on color perception. As a result, OpenAI-CLIP may be more suitable due to its more significant amount of training data that inherently includes more

versatile color variations. Thus, we use OpenAI-CLIP within our D-VST framework. On the other hand, while UNI offers strong pathological feature extraction, OpenAI-CLIP suffices for capturing tone information in D-VST. Hence, pathology foundation models trained on multi-stains like UNI are also excellent choices for the proposed D-VST framework. However, it is important to note that even when using PathCLIP as the tone encoder, our performance remains competitive with other methods in Table 1 of the main text.

Table A6: Performance comparison of using PathCLIP [76], UNI [10] and OpenAI-CLIP [63] as the tone encoder on the HE2IHC [61] dataset.

Tone encoder	DISTS↓	FID↓	KID↓	PSNR↑	SSIM↑
PathCLIP	0.190	40.01	0.0086	17.49±3.717	**0.411±0.136**
UNI	0.157	**33.13**	**0.0048**	18.03±3.842	0.404±0.130
OpenAI-CLIP	**0.154**	33.16	0.0055	**18.11±3.874**	0.407±0.136

Influence of hyper-parameter n for square split. The input images to be re-stained are divided into a grid of $n \times n$ squares for local frequency estimate. We vary n from 4 to 128 and show the FIDs for HE2IHC [61] in Table A7. When $n \in \{16, 32, 64\}$, the results are the best and stable, whereas n being too small or large deteriorates the performance. The empirical guideline is to select n properly so that each square contains enough pixels for a reliable frequency estimate but not too many pixels to remain a local estimate. We use $n = 32$ in our paper.

Table A7: Performance in FID with varying values for the hyper-parameter n on the HE2IHC [61] dataset.

n	4	8	16	32	64	128
FID↓	33.452	33.444	33.267	**33.162**	33.323	33.548

Ablation study on α and β. α and β are important hyper-parameters in our proposed frequency-aware adaptive patch sampling. Concretely, α controls the sampling probability discrepancy between low- and high-frequency regions, whereas β trades off between computational cost and image quality. Figure A7 presents the impact of varying α and β values on the model performance focused on FID, whose variations are more evident among the perception-oriented metrics.

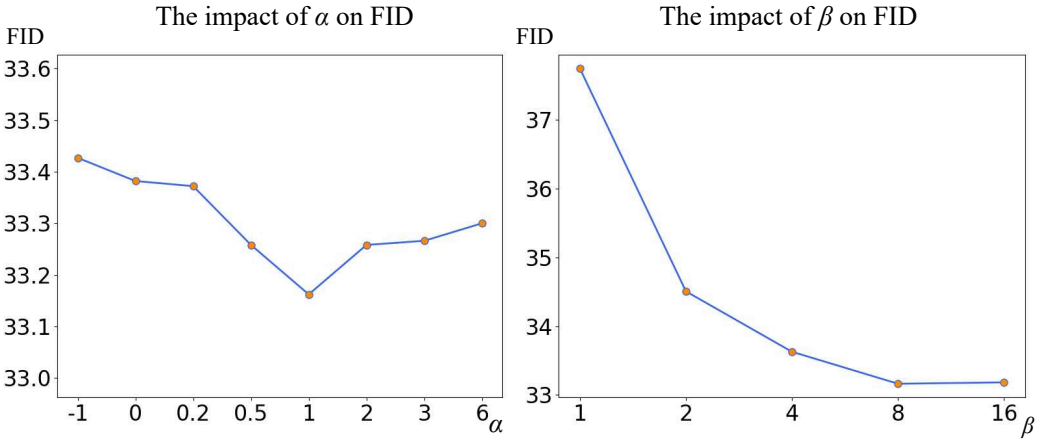

Figure A7: Performance plots (FID) of varying values of the hyper-parameters α (left) and β (right) on the HE2IHC [61] dataset. Note that the labels for the horizontal axes are indicative and do not strictly follow the actual intervals between values.

As shown in Figure A7 (left), the best performance occurs at $\alpha = 1$, while extreme values ($\alpha = 0$ or 6) degrade performance. Notably, $\alpha = 0$, which is effectively the random patch selection, yields the second worst performance. In addition, we experiment with sampling more patches in high-frequency

Table A8: Evaluation of various virtual staining methods on two external validation datasets (and virtual staining tasks). For all methods, the models trained for the corresponding tasks in the main text are directly used here for inference without any tuning.

Method	DISTS↓	FID↓	KID↓	PSNR↑	SSIM↑
Ext-HE2IHC [61]					
Pix2pix [37]	0.2923	147.02	0.1065	16.632±3.442	0.3250±0.126
Pix2pixHD [83]	0.2792	104.96	0.0597	16.361±3.390	0.3419±0.128
CycleGAN [100]	0.2823	149.61	0.1168	**16.918±3.484**	0.3410±0.117
[61]	0.2657	104.79	0.0796	15.708±2.234	0.3451±0.130
SynDiff [56]	0.3584	251.49	0.2500	16.700±3.356	0.3552±0.099
StainFuser [39]	0.2634	156.55	0.0932	16.131±3.336	0.3436±0.143
D-VST	**0.2628**	**88.287**	**0.0396**	16.713±3.598	**0.3601±0.140**
Ext-FFPE2HE [34]					
Pix2pix [37]	0.2066	44.919	0.0271	15.780±1.853	0.4580±0.076
Pix2pixHD [83]	0.1708	28.332	0.0139	17.726±1.457	0.4789±0.084
CycleGAN [100]	0.2913	99.343	0.0780	11.778±2.055	0.2774±0.125
[34]	0.1908	51.966	0.0342	**18.975±1.404**	**0.5932±0.073**
SynDiff [56]	0.3420	154.37	0.1344	6.8810±2.005	0.2219±0.114
StainFuser [39]	0.2153	71.887	0.0408	12.538±1.051	0.2326±0.093
D-VST	**0.1521**	**27.488**	**0.0108**	17.179±1.036	0.4591±0.075

regions by setting $\alpha = -1$. The performance is worse than that of random sampling ($\alpha = 0$) and our low-frequency-preferred sampling ($\alpha = 1$), validating that mosaic artifacts impact low-frequency regions more. We use $\alpha = 1$ for experiment comparison with other methods in the main text.

As for β, we study its impact with α fixed to 1. Figure A7 (right) shows that increasing the number of sampled patches rapidly improves the performance, which is reasonable, until the saturation at $\beta = 16$—with a similar performance to $\beta = 8$. Considering that $\beta = 16$ doubles the amount of computation of $\beta = 8$, we set $\beta = 8$ for performance comparison in the main text.

Validate sampling probability design. To validate our design of sampling probabilities in Eqn. (2), we use the medium l of images as the threshold for low- and high-frequency patches. It turns out that our method samples 69% of patches in low-frequency squares versus 31% in high-frequency ones. These numbers indicate that the sampling probabilities work as designed.

External validation. To further evaluate the generalizability of our method, we conduct external validation on two datasets: (1) the external test set from [61], comprising 285 HE-IHC image pairs stained with SP3 [14] or CB11 [62] antibodies (Ext-HE2IHC); and (2) the external test set from [34], containing 1,398 FFPE-HE image pairs (Ext-FFPE2HE). It is worth noting that for the external validation, we directly use the models trained for the corresponding tasks in the main text without further tuning. As shown in Table A8, D-VST again achieves the best performance for the three perception-oriented metrics (DISTS, FID, and KID) on both external test datasets, and competitive performance for PSNR and SSIM (ranking top one to top three among all compared methods). These results are consistent with those presented in the main text, demonstrating D-VST's strong generalizability in cross-dye virtual staining of histopathology images.

Additional comparison with existing methods. In this section, we compare our D-VST with two additional state-of-the-art approaches to virtual staining, PSRVS [98] and DeepLIIF [23], on the HE2IHC [61] dataset. The former belongs to diffusion-based models, and the latter to GAN-based. The results are shown in Table A9. With the due caution that these two methods may not be fully optimized for this task, we can see that D-VST substantially outperforms PSRVS and DeepLIIF in DISTS, FID, and KID, and is comparable in PSNR and SSIM.

Additional qualitative results. We present additional qualitative tone control results on HE2IHC and FFPE2HE in Figure A9, under settings consistent with Table 1 and Figure 5 of the main text. We also show qualitative visual comparisons with other methods in Figure A10 and Figure A11 under the same settings. Figure A12 displays more qualitative comparisons between different sampling strategies

Table A9: Additional performance comparison with PSRVS [98] and DeepLIIF [23] on the HE2IHC dataset.

Method	DISTS↓	FID↓	KID↓	PSNR↑	SSIM↑
D-VST (ours)	**0.154**	**33.16**	**0.0055**	**18.11±3.874**	0.407±0.136
PSRVS	0.422	230.3	0.2283	17.94±4.193	0.396±0.128
DeepLiiF	0.305	140.3	0.1379	17.96±3.713	**0.418±0.129**

for zero-shot virtual staining of large histopathology images at the resolution of 2048×2048 pixels. These results qualitatively demonstrate that D-VST can (1) precisely control the tones of the cross-dye virtually stained histopathology images without pathology leakage, (2) generate high-quality large histopathology images in an efficient manner, and (3) support virtual staining of ultra-high-resolution WSIs, validating the versatile generative capacity and generalization ability of D-VST.

WSI validation. As far as we know, no suitable WSI dataset is currently available for reliable quantitative evaluation at scale. Therefore, we mainly show the qualitative results in Figure A13, Figure A14, and Figure A15 for an observational study. Figure A13 shows ultra-high-resolution (16,384×16,384 pixels) HE2IHC virtual staining examples. Lastly, Figure A14 and Figure A15 showcase virtual staining results of WSIs (16,000×15,096 and 40,000×32,496 pixels) from Ext-FFPE2HE, each conditioned with two different target tones. We have asked two board-certified pathologists to blindly rank the WSIs virtually stained by our D-VST, a representative GAN-based method (pix2pixHD [83]), and two diffusion-based methods (SynDiff [56] and StainFuser [39]), considering image quality and pathology correctness. The evaluated WSIs include two HE2IHC and two FFPE2HE images, shown in in Figure A13, Figure A14, and Figure A15, respectively. The mean ranking is: D-VST (1.0), pix2pixHD (2.5), StainFuser(3.0), and SynDiff (3.5), demonstrating D-VST's superior virtual staining quality.

As for timing, we record the generation times for the WSI in Figure 1 of the main text (16,384×16,384 pixels; HE to IHC) using a representative GAN-based method (pix2pixHD [83]), two diffusion-based methods (StainFuser [39] and SynDiff [56]), and our D-VST. To avoid the unwanted mosaic artifacts, we implement the MultiDiffusion [3] sampling strategy for the compared methods, whereas D-VST uses its adaptive sampling strategy. As shown in Table A10, D-VST is orders of magnitude faster than the other diffusion-based methods and comparable to the GAN-based approach, highlighting its efficiency advantage for large WSI virtual staining over existing diffusion-based methods.

Table A10: Runtime comparison of Pix2pixHD [83], StainFuser [39], SynDiff [56], and D-VST for generating a WSI of 16,384×16,384 pixels.

Method	Pix2pixHD	StainFuser	SynDiff	D-VST (ours)
Time (second)	7,127	840,499	172,032	8,862

Influence of fine-tuning variational autoencoder (VAE). D-VST inherits the Stable Diffusion (SD) VAE [67] from PixArt-α [9], as it adopts the Diffusion Transformer (DiT) from PixArt-α as the denoising network. Our empirical evidence, both quantitative and qualitative, shows that D-VST with the SD VAE outperforms various GAN- and diffusion-based approaches on both virtual staining and downstream classification tasks. Thus, we argue that the SD VAE suffices for encoding and decoding of pathology images in D-VST, despite being pretrained on natural images and probably not being the optimal choice. We attribute this to SD VAE's strong encoding/decoding ability from large-scale, diverse training, and DiT's high compatibility with it. Notably, other works (e.g., HistDiST [25], StainFuser [39]) also successfully used VAEs pretrained on natural images for histopathology virtual staining.

We also experimented with fine-tuning the SD VAE on the HE2IHC dataset, but observed mixed results: similar PSNR/SSIM but worse DISTS, FID, and KID. We visualize the reconstruction error distributions in Figure A8. The reconstructed images show blurred cell membranes in lesion regions. As shown in the MSE maps, these lesion regions exhibit the highest intensity, indicating that the reconstructed images deviate most from the ground truth in these areas. We speculate that this

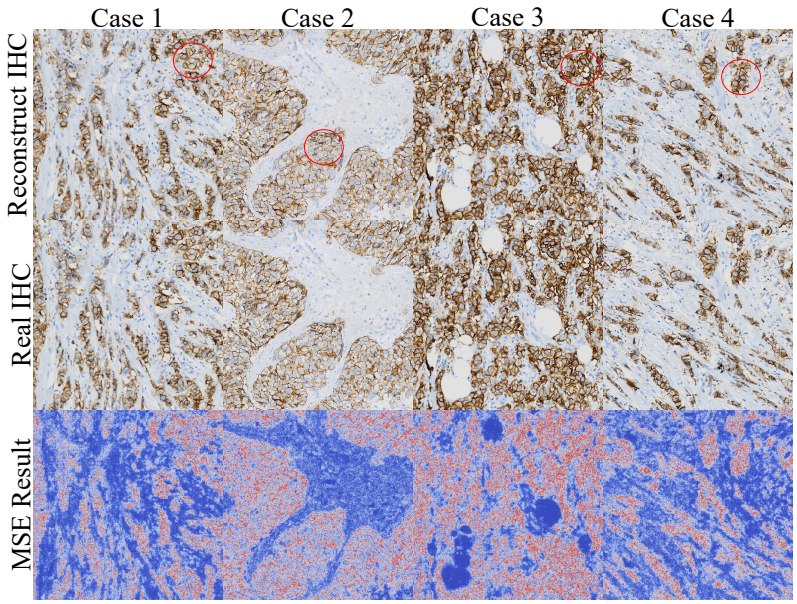

Figure A8: *Reconstruct IHC* denotes the generated images produced with the fine-tuned VAE, whereas *Real IHC* represents the corresponding ground-truth images. *MSE Result* refers to the pixel-wise mean squared error (MSE) maps between the reconstructed and real images, where regions with higher temperature indicate larger discrepancies. The red ellipses outline some areas where the reconstructed cell membranes in lesion regions appear noticeably blurred.

degradation is due to the limited training data, which prevented the VAE from learning pathological structures effectively; however, further validation with more data is needed.

Impact of tone control image selection on the downstream task. To evaluate the robustness of D-VST to tone control images, we conduct an additional set of experiments in which only a single IHC conditioning patch is used. Specifically, we randomly select one IHC image from the sub-test set as the sole tone conditioning patch and repeat the experiment three times with different selections, reporting the averaged results in Table A11. Otherwise, these experiments follow the same setting as Table 3 in the main text. As we can see, the performance with a single IHC conditioning patch is comparable to that obtained using multiple patches (originally reported in Table 3). This demonstrates that D-VST's downstream task performance is robust to the number of tone conditioning images, which we attribute to its effective disentanglement of tone and pathological conditions.

Table A11: Impact of tone control image selection on the downstream task.

No. tone patches	ACC	F1	Precision	Recall
Multiple (original)	0.9417	0.9430	0.9470	0.9388
Single (avg. over 3 runs)	0.9415$\pm$0.0068	0.9427$\pm$0.0064	0.9483$\pm$0.0050	0.9416$\pm$0.0068

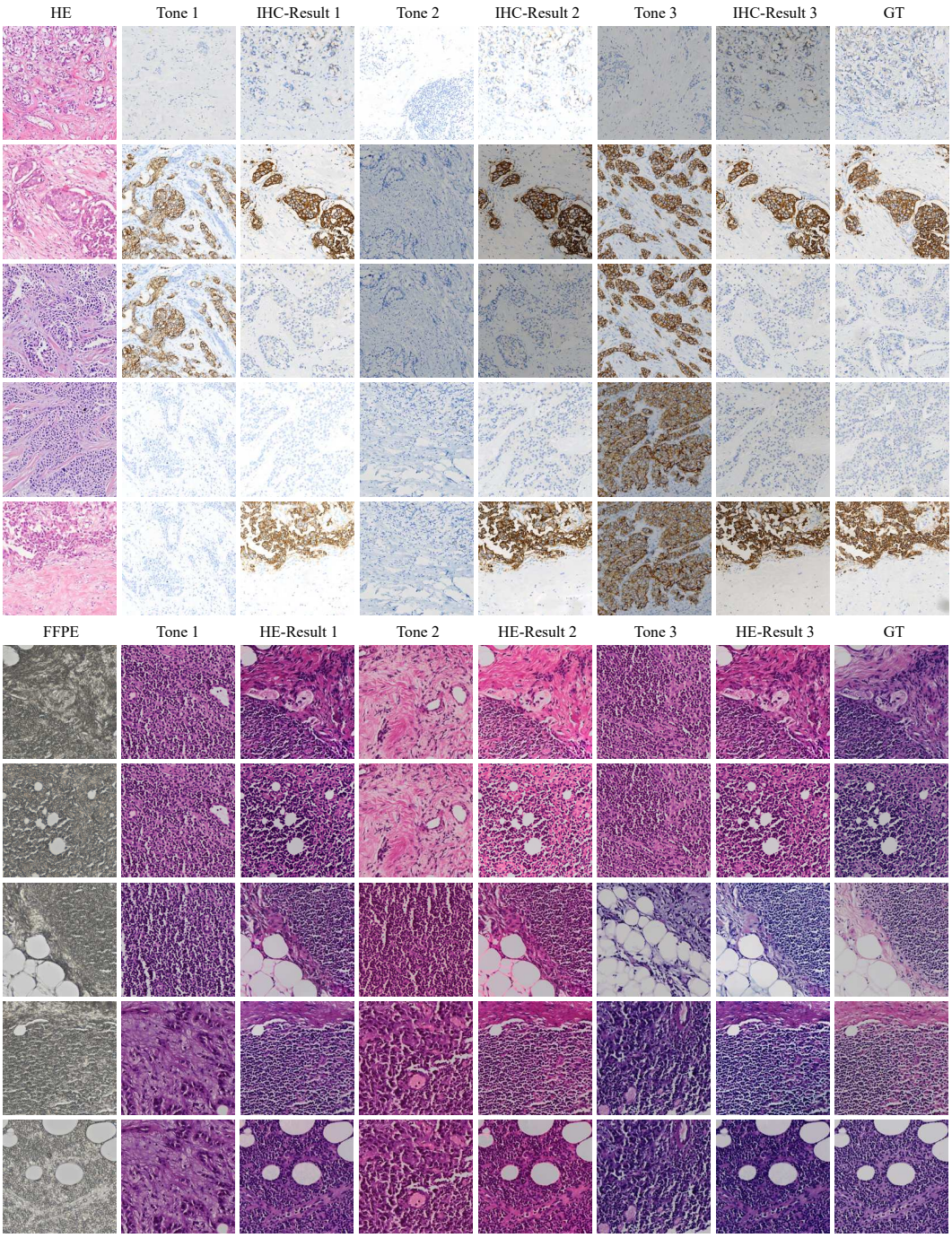

Figure A9: Additional qualitative results for tone control. Top: HE to IHC; and bottom: FFPE to HE. The images are virtually stained at the resolution of 1024×1024 pixels. GT: ground truth.

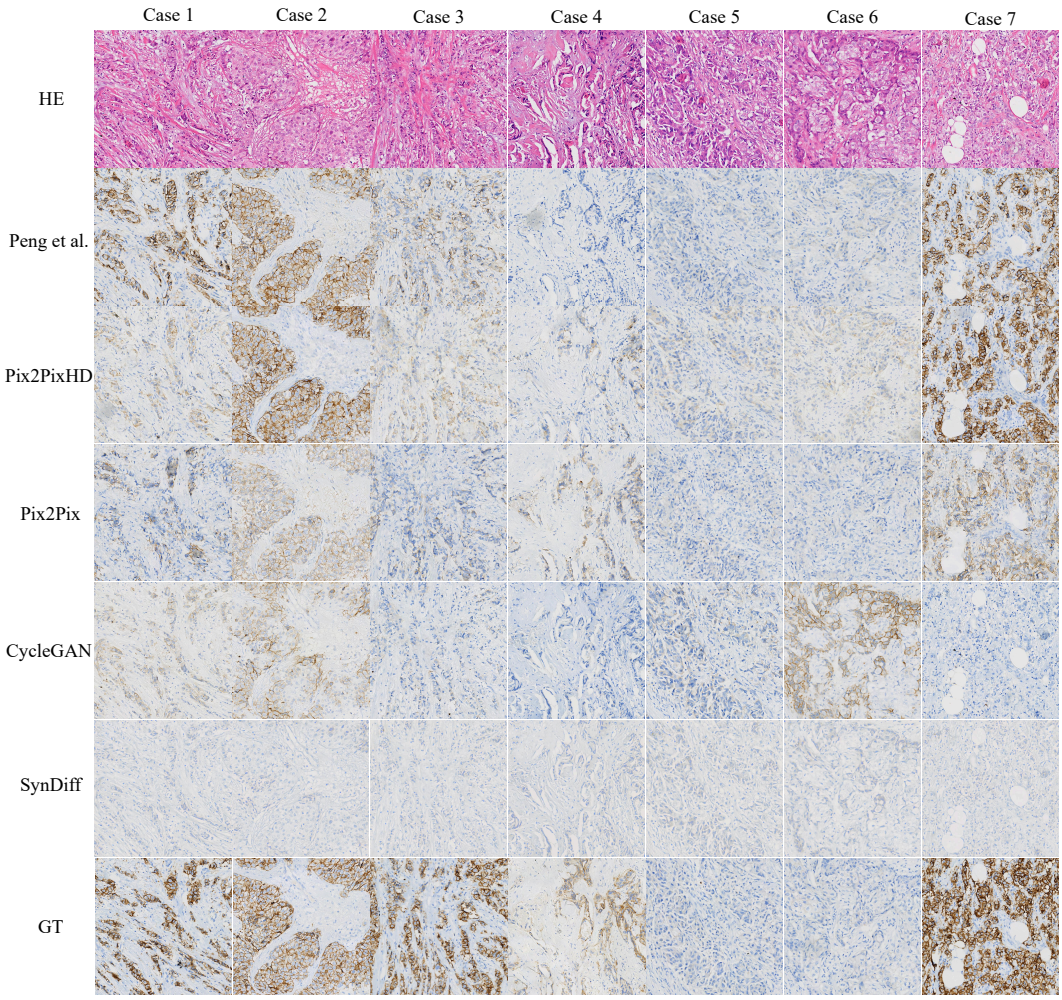

Figure A10: Qualitative comparison of HE to IHC virtual staining results by various methods that cannot control the tone of the virtually stained images, including Peng et al. [61], Pix2PixHD [83], Pix2Pix [37], CycleGAN [100], and SynDiff [56]. Note that the corresponding results by a few methods that control the tone of the re-stained images through image-based conditioning (including ours) are presented in Figure A11 for comparison. The images are virtually stained at the resolution of 1024×1024 pixels. GT: ground truth. By comparing the virtually stained images with the GT, we can observe clear Type I (false positive, meaning hallucinated cancerous status) and Type II (false negative, meaning hallucinated cancer-free status) errors for these methods.

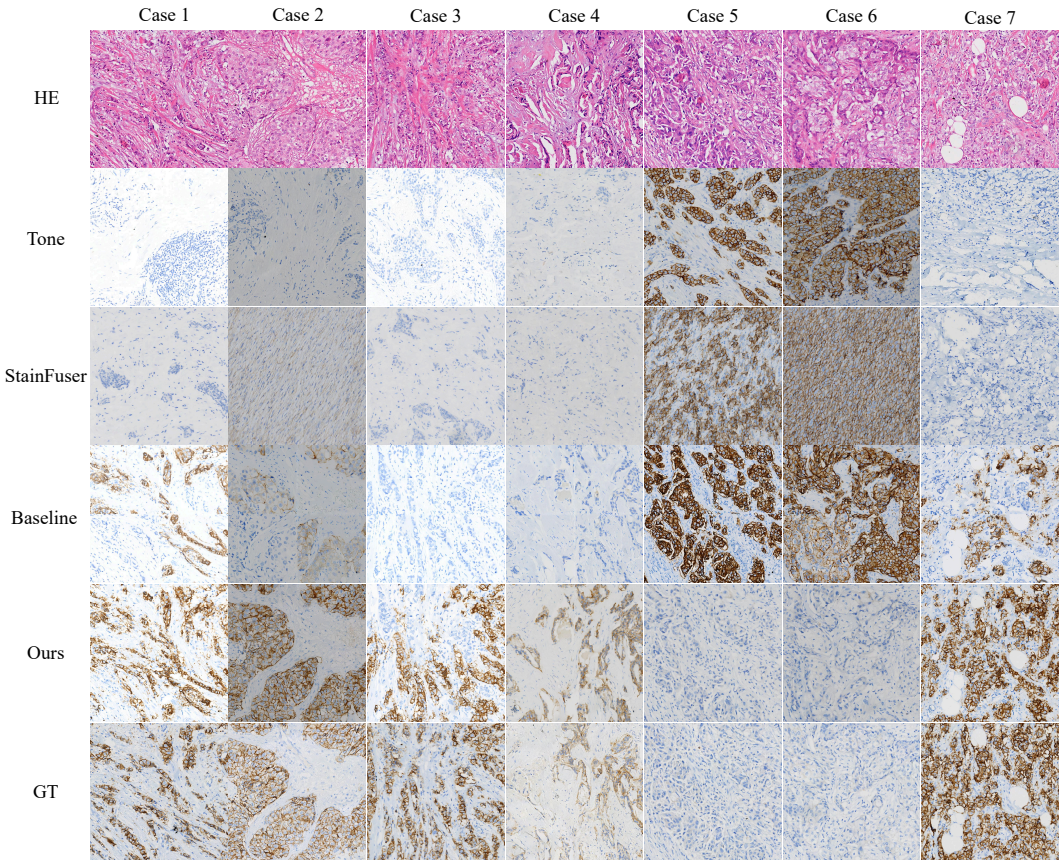

Figure A11: Qualitative comparison with other methods for HE to IHC virtual staining with tone control, including StainFuser [39], baseline (training the same network in Figure 3 (left) as our D-VST but without the proposed two-stage condition-decoupling curriculum learning). The "Tone" row displays the tone-conditioning images. Note that the corresponding results by methods that cannot control the tone of the re-stained images are shown in Figure A10 for comparison. The images are virtually stained at the resolution of 1024×1024 pixels. GT: ground truth. By comparing the virtually stained images with the GT, we can observe clear pathology leakages of Type I (false positive, meaning hallucinated cancerous status) and Type II (false negative, meaning hallucinated cancer-free status) errors for the compared methods. In contrast, the results of our D-VST align closely with the pathological statuses and distributions of the GT, while accurately reflecting the tones of the conditioning images.

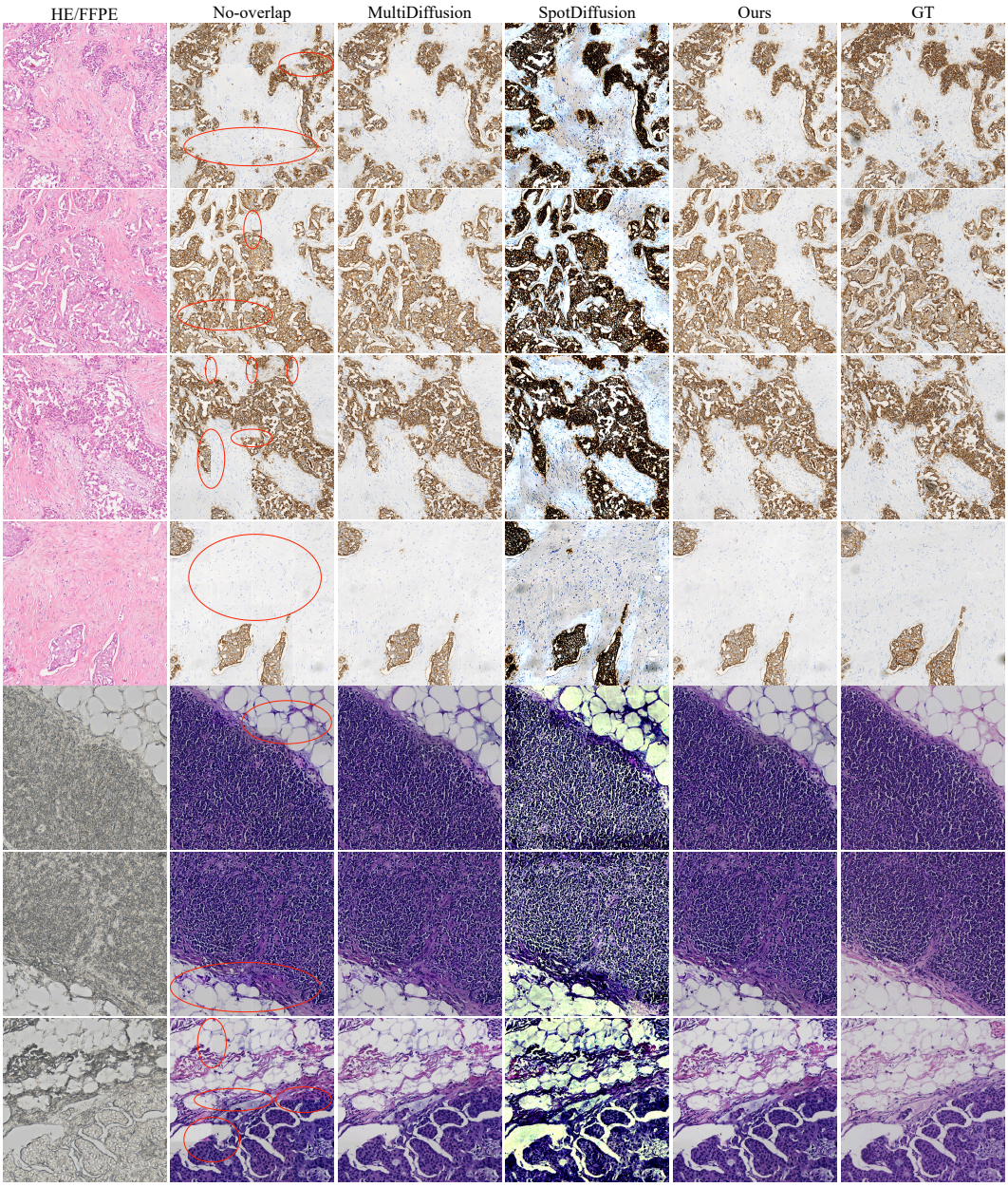

Figure A12: Comparison of sampling strategies for zero-shot virtual staining of large histopathology images (2048×2048 pixels). The red ellipses outline regions where the mosaic artifacts are prominent (best viewed zoomed in). GT: Ground truth.

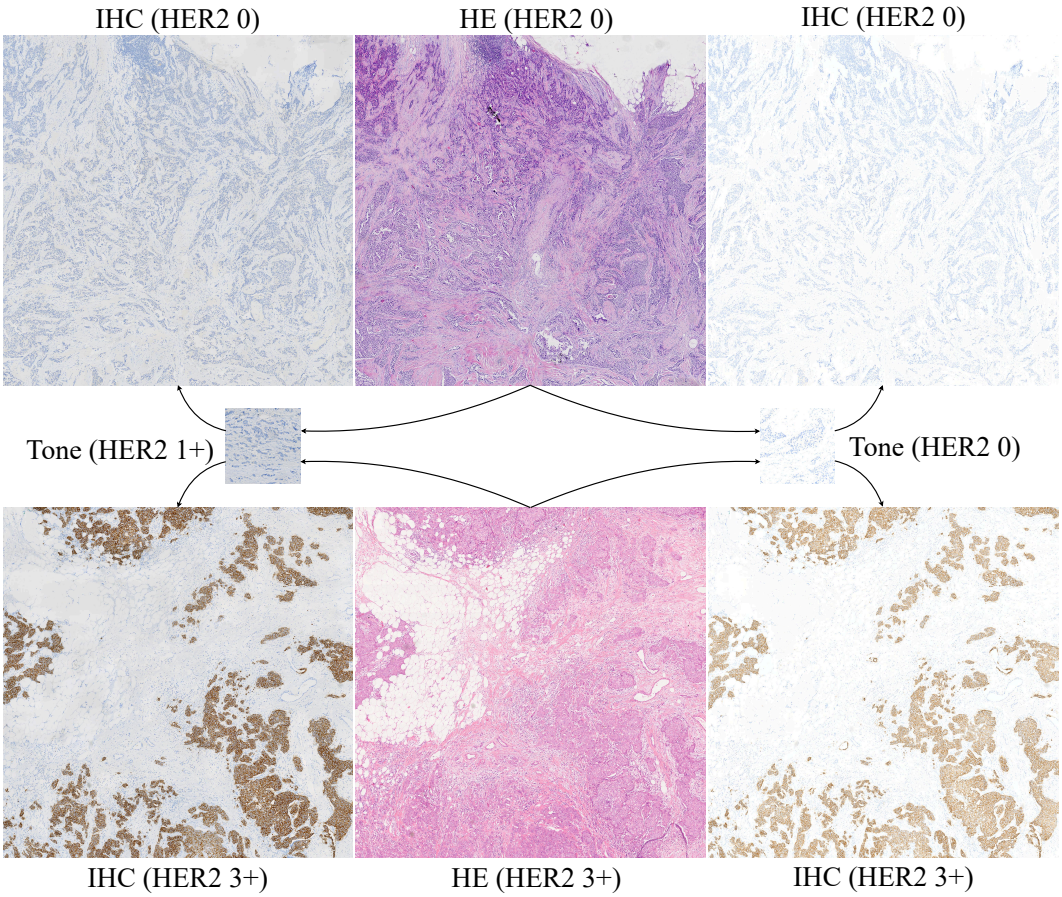

Figure A13: Examples of ultra-high-resolution (16,384×16,384 pixels) HE2IHC virtual staining. The central column shows the source-domain HE images, whereas the left and right columns show the virtually stained IHC images conditioned on two different tone-control images from the target domain. The top and bottom rows show two examples. HER2 scores: 0: no cancerous lesion, and 1+, 2+, and 3+: increasing severity of cancerous lesions.

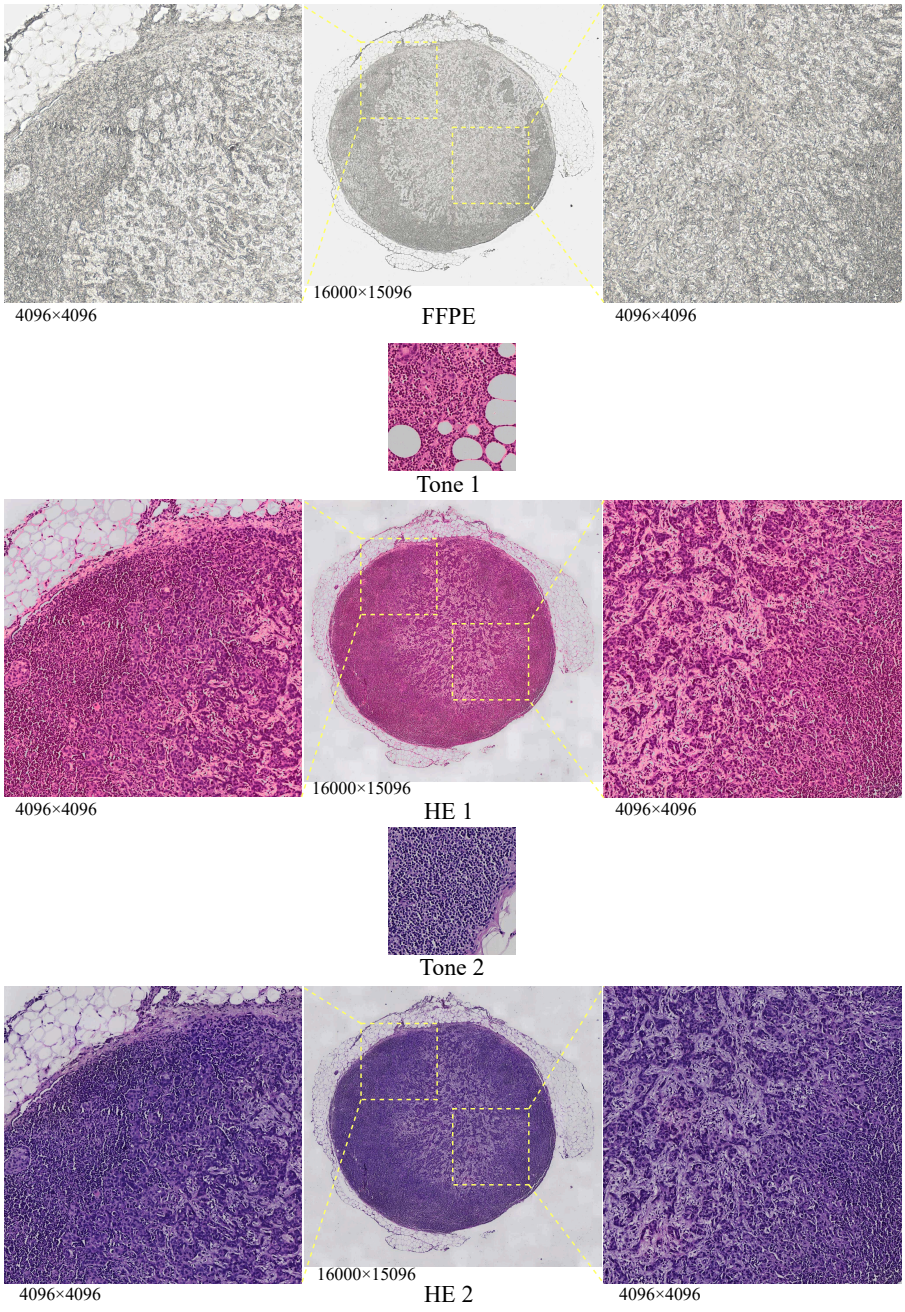

Figure A14: Virtual staining of an FFPE WSI in Ext-FFPE2HE, illustrated with two tone-conditioning images (Tone 1 and Tone 2) from the target domain. The virtual staining results corresponding to Tone 1 and Tone 2 are HE 1 and HE 2, respectively.

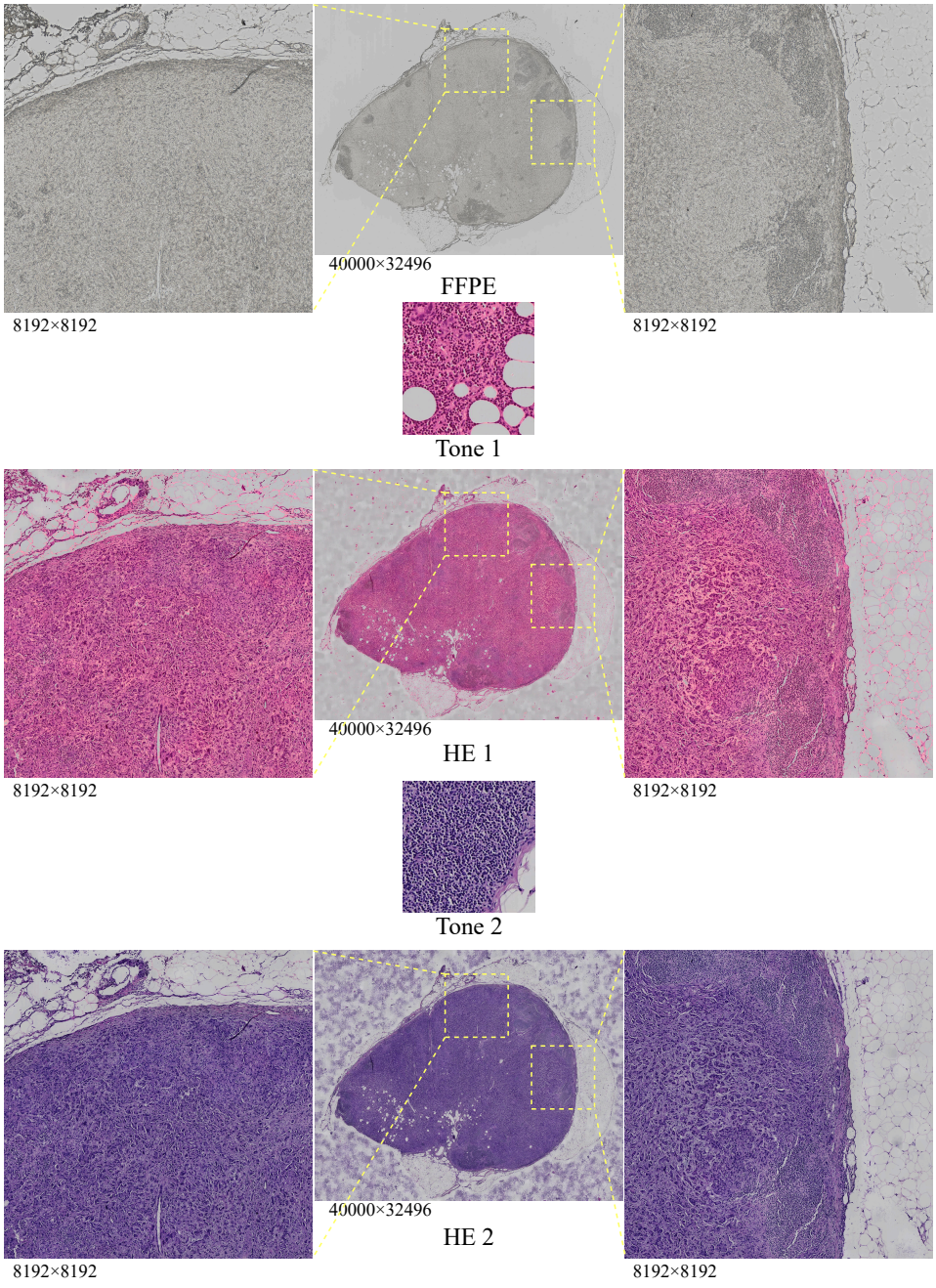

Figure A15: Virtual staining of an FFPE WSI in Ext-FFPE2HE, illustrated with two tone-conditioning images (Tone 1 and Tone 2) from the target domain. The virtual staining results corresponding to Tone 1 and Tone 2 are HE 1 and HE 2, respectively.