MAGICSTAIN: HIGH-FIDELITY PATHOLOGY IMAGE VIRTUAL STAINING VIA GUIDED SINGLE-STEP DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Virtual staining leverages computational methods to generate different stain styles from a pathology image of a chemically stained tissue section, offering a cost-effective alternative to chemical multiple staining. Despite extensive research based on generative adversarial networks (GANs) and diffusion models, achieving high-fidelity, high-quality, and computationally efficient virtual staining remains a significant challenge. While diffusion approaches typically produce more photo-realistic images than GAN counterparts through multi-step sampling, this comes at the cost of high computational overhead and inference latency. This paper proposes MagicStain, a novel single-step diffusion model tailored for generating high-resolution virtual stains. Specifically, we adapt a single-step diffusion model to enable efficient virtual staining. By introducing pathology priors from a pathology foundation model and integrating pathology- and structure-consistency losses on both the original images and the hematoxylin channel, MagicStain achieves high-fidelity, high-quality generations. To address the limitations of single-step diffusion models in high-resolution virtual staining, we further propose a two-stage progressive training strategy that enables high-resolution adaptation at low training cost. Extensive experiments on *eight* virtual staining datasets, each involving translation between different staining dyes/biomarkers, demonstrate the superiority of MagicStain in fidelity, visual quality, and computational efficiency compared to existing methods. Our code and trained models will be publicly available.

1 INTRODUCTION

Histochemical staining is widely used to distinguish various tissues by colorization, thus facilitating pathological diagnosis De Cuyper et al. (2020); Iqbal & Iqbal (2014); Stack et al. (2014). Different dyes manifest different colors in stained tissue and yield complementary information. For instance, hematoxylin-eosin (H&E) highlights cellular morphology, while immunohistochemistry (IHC) reveals protein-specific expression critical for tumor identification and cancer prognosis Bai et al. (2023). However, current chemical protocols allow only a single stain per tissue section, necessitating additional sections for multiple stains. This not only increases the consumption of often-limited clinical tissue samples but also incurs substantial time and chemical costs. As a result, performing multiple stains is resource-intensive, laborious, and expensive Bai et al. (2023).

Virtual staining Bai et al. (2023); Khan et al. (2023) offers a promising and cost-effective alternative to conventional chemical staining for generating multiple stains. Being essentially an image-to-image (I2I) translation task, it digitally "translates" chemically stained pathology images using computational methods, converting images stained with one dye (the source domain) to new images that resemble those stained with another (the target domain), e.g., H&E to IHC. Ideally, the virtually stained image should be of high quality (e.g., clear, sharp, and exhibiting the same brightness and color style as the target domain) and high fidelity (accurately representing the source image's authentic content and pathology).

Researchers have explored generative adversarial networks (GANs) Goodfellow et al. (2014) and diffusion models Ho et al. (2020) for virtual staining. GAN-based methods utilize image style transfer baselines (e.g., pix2pix Isola et al. (2017), CycleGAN Zhu et al. (2017)) to translate the staining style

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

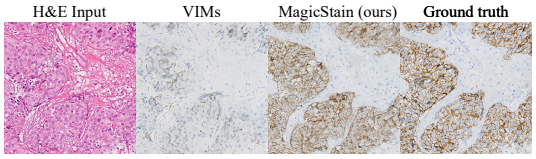

Figure 1: Existing single-step pathology image diffusion model VIMs Dubey et al. (2024); Parmar et al. (2024) fails to correctly transfer the pathology of source images to the target domain: Compared to the ground truth, although the overall distribution and structure of the blue dots (representing nuclei) in the generated IHC image appear somewhat realistic, the image lacks the brownish components indicating cancerous regions, likely due to a lack of domain expertise. Our prior-driven MagicStain synthesizes the correct pathology while achieving high structural accuracy.

of source-domain images to that of the target domains. They often incorporate additional modules or loss functions to explicitly enforce structural or pathological consistency, leading to improved fidelity Peng et al. (2024). Diffusion models offer a higher quality ceiling and have been applied to virtual staining Großkopf et al. (2025); Jewsbury et al. (2024); Yellapragada et al. (2025), but suffer from substantial inference latency and computational costs due to multi-step sampling.

To expedite inference and improve the practical applicability of diffusion models, various methods distill slow, many-step teacher models into faster, few-step students Luo et al. (2023); Xu et al. (2024); Yin et al. (2024). SD-Turbo Sauer et al. (2024) distills Stable Diffusion Rombach et al. (2022) into a one-step model via adversarial training, and pix2pix-Turbo Parmar et al. (2024) extends it to structure-preserving I2I translation by modifying the input conditioning of SD-Turbo Sauer et al. (2024) and fine-tuning it. VIMs Dubey et al. (2024) first applies the training paradigm of pix2pix-Turbo Parmar et al. (2024) directly to virtual staining at the 512 resolution, providing an encouraging proof of concept.

However, directly fine-tuning SD-Turbo Sauer et al. (2024) for virtual staining like VIMs exposes a fundamental problem: VIMs suffers a severe performance drop at the 1024 resolution—the standard benchmark resolution on which GAN-based virtual staining methods are commonly evaluated Bian et al. (2024); Klöckner et al. (2025); Li et al. (2023a); Liu et al. (2022); Peng et al. (2026). Compared to the ground truth (Figure 1), although the overall structure of the blue dots (representing nuclei) in the IHC images generated by VIMs appears somewhat realistic, these images lack the brownish components that indicate cancerous regions (more results in Figure 3, and Tables 1 and 4).

We identify three compounding issues for this failure. First, unlike pix2pix-Turbo Parmar et al. (2024), which only needs to fill in visually plausible textures from abstract structural signals (e.g., edge maps or sketches), virtual staining requires simultaneously preserving source-domain morphology and learning an implicit mapping from morphology to pathological semantics—a level of domain-specific reasoning that goes beyond texture synthesis. Second, the supervision in pix2pix-Turbo Parmar et al. (2024) targets natural image generation (and its baseline model, SD-Turbo Sauer et al. (2024), is pretrained entirely on natural images), resulting in insufficient pathological knowledge for virtual staining. VIMs inherits this limitation by naively adopting pix2pix-Turbo (we verify in Tables 2 and 3 that introducing pathology priors and supervision yields consistent improvements). Third, these issues are further amplified at higher resolutions (e.g., 1024×1024 pixels), where training becomes substantially harder—both for accurate pathological style transfer and for learning fine-grained morphological structures (see the resolution ablation in Table 3).

This work presents MagicStain, an approach to high-fidelity (in terms of correct structure and pathology) and efficient single-step, high-resolution virtual staining of pathology images. We propose a series of targeted solutions to address the issues above. To tackle the lack of domain knowledge and explicit pathology constraints, we design a pathology-prior-aided pipeline centered on PathCLIP Radford et al. (2021); Sun et al. (2024b), a pathology foundation model (PFM). Specifically, we use PathCLIP both as a prior extractor—injecting pathology priors from the source image into the denoising network—and as a supervision signal, enforcing alignment between the PathCLIP representations of the generated and ground-truth target-domain images. This closes a self-consistent loop: the same pathology expert that guides generation also judges its fidelity.

To further generalize to the higher resolution of 1024×1024 pixels commonly used in virtual staining benchmarks, we propose a progressive training strategy. We first train the model on low-resolution images for rapid convergence, then fine-tune it on high-resolution images using low-rank adaptation (LoRA) Hu et al. (2022). Our experiments demonstrate that this strategy enables fast generalization from low to high resolution, and achieves better performance than directly training a high-resolution model from scratch or applying a low-resolution model to high-resolution generation. We also impose perceptual and reconstruction losses on the hematoxylin (H) channel throughout both training stages, leveraging the medical prior that the H channel highlights nuclei in both H&E and IHC images. These losses provide finer-grained morphological supervision that is otherwise diluted at high resolution when losses are applied on the entire image (cf. the ablation in Table 3 for their effectiveness).

In summary, our contributions are as follows:

(1) We introduce MagicStain, a simple yet effective single-step diffusion model that enables high-fidelity and high-resolution virtual staining of pathology images.

(2) We seamlessly incorporate domain-specific modules and loss functions driven by medical priors, ensuring accurate translation of cellular structures and pathology.

(3) We propose a progressive training strategy for high-resolution adaptation. To our knowledge, MagicStain is the first single-step diffusion model that synthesizes virtual stains at a resolution of 1024×1024 pixels.

(4) Last but not least, we conduct extensive experiments on *eight* virtual staining datasets to validate MagicStain against state-of-the-art methods. We also assess a downstream task and perform ablation studies on MagicStain.

2 Related Work

GAN-based Virtual Staining. GAN-based methods have been widely adopted for virtual staining: H&E to IHC Bian et al. (2024); Klöckner et al. (2025); Li et al. (2023a); Liu et al. (2022); Peng et al. (2026), H&E synthesis from other modalities Asaf et al. (2024); Hu et al. (2024); Khan et al. (2023); Li et al. (2024b); Rahman Saurav et al. (2025); Rivenson et al. (2019), and broader cross-modality conversions Lee et al. (2026); Liu et al. (2024); Ma et al. (2023); Xu et al. (2025b). Several methods also explored predicting pathology-relevant representations directly, without generating the full target image Zhu et al. (2026), or synthesizing panel proteins in multiplexed imaging Możejko et al. (2026). To improve pathological fidelity, many GAN-based approaches incorporated domain-informed supervision, including classification labels Jia et al. (2025); Li et al. (2024a;b); Liu et al. (2021); Xu et al. (2019), optical density maps Chen et al. (2026; 2024a); Peng et al. (2026), and diaminobenzidine (DAB) channel supervision Ma et al. (2026); Peng et al. (2024); Saurav et al. (2026); Wei et al. (2024); Xu et al. (2025a) to enforce accurate transfer of pathology-specific signals. For structural fidelity, prior work leveraged H-channel similarity Wei et al. (2024), topology-aware constraints Jiang et al. (2026), local variance matching Lee et al. (2026), structural priors from third-party segmentation tools Hu et al. (2024); Peng et al. (2024); Wang et al. (2024a), embeddings from self-supervised pathology models Saurav et al. (2026); Yang et al. (2025), frequency-domain supervision Zhang et al. (2024), and a variety of auxiliary structural losses Dubey et al. (2023); Pati et al. (2024); Sloboda et al. (2023); Wang et al. (2024b). Content-style decoupling has also been explored to disentangle morphological structure from staining appearance Wang et al. (2025); Yuan et al. (2025). Despite notable progress, GAN-based methods still fall short in generating images with high perceptual quality Rombach et al. (2022), motivating the adoption of diffusion models.

Diffusion-based Virtual Staining. Recent advances in diffusion models Rombach et al. (2022); Zhang et al. (2023) have demonstrated superior image synthesis quality over GANs, spurring their application to virtual staining Ho et al. (2025); Jewsbury et al. (2024); Kataria et al. (2025); Tian et al. (2026); Xiong et al. (2025); Yan et al. (2025); Zheng et al. (2025). Chen et al. (2025); Großkopf et al. (2025); Yellapragada et al. (2025) incorporated third-party models for pathology and structure priors. He et al. (2024b); Liu et al. (2025) employed asymmetric attention on pathological regions or frequency-domain information to preserve structural details. Kataria et al. (2024) generated IHC images and segmentation masks in parallel via dual diffusion models to impose morphological supervision. However, all these many-step diffusion methods suffer from high inference latency,

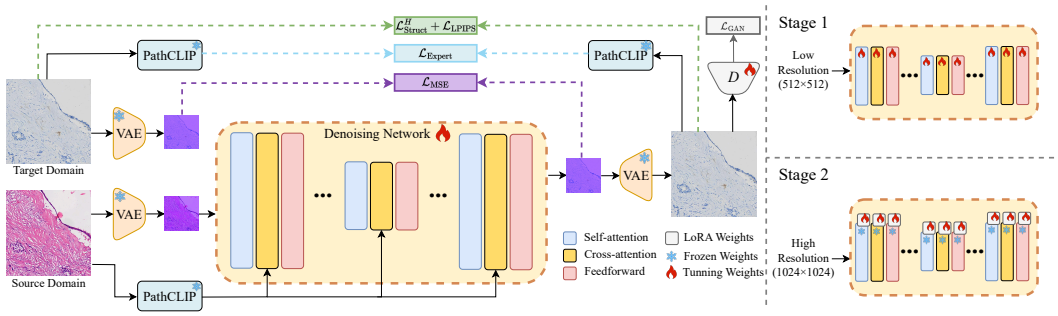

Figure 2: Left: **Framework of the proposed MagicStain**. MagicStain adapts SD-Turbo, a text-to-image one-step diffusion model, for single-step pathology image virtual staining. Additionally, it incorporates PathCLIP, a specialized pathology foundation model (PFM) that offers expertise in pathology images. The small, purple-colored pathology images are only for schematic illustration of latent-space representations (i.e., output of the VAE encoder and input of the VAE decoder). Right: **Progressive training for high-resolution adaptation.** Stage 1 trains all parameters using 512×512-pixel images. Stage 2 updates only the LoRA weights using 1024×1024-pixel images.

and—due to their iterative denoising process Song et al. (2020)—it is difficult to impose direct pathology and structural constraints on the final output, a strategy proven effective in GAN-based approaches.

To accelerate diffusion model inference, various distillation strategies have been proposed, including consistency training Luo et al. (2023); Song et al. (2023), adversarial distillation Parmar et al. (2024); Sauer et al. (2024); Xu et al. (2024), variational score distillation Yin et al. (2024), and rectified flow Liu et al. (2023). Model quantization He et al. (2024a); Li et al. (2023b;c) and pruning Han et al. (2015) offer complementary acceleration avenues. For virtual staining, VIMs Dubey et al. (2024) pioneered the use of pix2pix-Turbo Parmar et al. (2024), a single-step image-to-image translation framework, achieving proof-of-concept results at the resolution of 512×512 pixels. However, VIMs did not incorporate any pathology-domain knowledge, and single-step diffusion models are particularly susceptible to this gap: unlike many-step models that can iteratively refine outputs under guidance, a single-step model must perform the entire semantic mapping in a single forward pass, making the absence of pathology priors especially detrimental. In contrast, MagicStain addresses both issues via a single-step diffusion model, augmented with a specialist PFM for pathology expertise and a two-stage training strategy for effective high-resolution synthesis of 1024×1024-pixel images.

3 Method

3.1 Overview

Our target is to translate images from a source domain $\mathcal{X} \subset \mathbb{R}^{H \times W \times 3}$ (e.g., H&E) to some desired target domain $\mathcal{Y} \subset \mathbb{R}^{H \times W \times 3}$ (e.g., IHC), where H and W are the height and width of the images, by learning a translation function $\hat{y} = f_G(x) : \mathcal{X} \to \mathcal{Y}$ on a paired dataset $\mathcal{D} = \{(x, y) | x \in \mathcal{X}, y \in \mathcal{Y}\}$. The generated image $\hat{y}$ should be as close as possible to the real target-domain image y in terms of not only color, brightness, and contrast, but also tissue structure and pathology status.

The framework of our method, MagicStain, is shown in Figure 2. It adapts a text-to-image one-step diffusion model, SD-Turbo (Sauer et al., 2024), which includes the denoising network and the variational autoencoder (VAE) (Kingma & Welling, 2022). Parmar (Parmar et al., 2024) observed that standard diffusion adapters like ControlNet (Zhang et al., 2023) were less effective for one-step I2I translation that requires detailed structure preservation, as the noise map influences the output structure. Therefore, we follow pix2pix-Turbo (Parmar et al., 2024) to directly feed the latent representation of the source image encoded by the VAE encoder to the denoising network, without the noise map or additional branches in conventional many-step diffusion models. The latent representation is forwarded through the denoising network only once, whose output is decoded by the VAE decoder to produce a re-stained pathology image. In addition, adversarial learning is incorporated to ensure the distribution of generated images is close to that of the target domain.

For effective and efficient single-step virtual staining of pathology images, MagicStain proposes three innovative designs. First, it incorporates PathCLIP, a pathology foundation model (PFM), as an "expert" assistant. On the one hand, PathCLIP encodes source images into pathology-prior embeddings to guide the I2I translation via multi-head attention. On the other hand, it extracts embeddings from the generated and real target-domain images, with a loss enforcing their similarity. Second, based on the medical prior that the H channel reflects nuclear morphology in both H&E and IHC images, we impose additional structural and semantic consistency on the H channel to enhance structure and pathology fidelity. Third, we adopt a progressive training strategy for high-resolution adaptation, ensuring high-resolution performance while maintaining low training cost and avoiding generalization degradation. Our empirical results validate the efficacy of MagicStain's designs.

3.2 Incorporate Pathology Expert Prior via Specialist PFM

Although some existing I2I translation methods have introduced various pathology priors for virtual staining, they still lack in-depth, comprehensive expertise in pathology. Recently, PFMs trained on large-scale pathology image-text pairs have demonstrated strong capabilities in extracting histopathological features critical to clinically relevant downstream tasks (Sun et al., 2024a;b). Inspired by their effectiveness, we adopt PathCLIP (Sun et al., 2024b), a PFM trained on 207K high-quality pathology image-caption pairs via contrastive learning, to provide expert-level pathology prior in our MagicStain framework at both the input and output ends.

At the input end, PathCLIP serves as an expert-level assistant, providing extensive prior knowledge about the source image to support clinically correct virtual staining—akin to a seasoned pathologist. Specifically, we utilize the image encoder of PathCLIP to encode the source images into pathology-aware embeddings. The embeddings are used in place of the conditioning textual embeddings (i.e., the controlling prompt in the original SD-Turbo architecture), serving as the keys and values for multi-head cross-attention in the denoising network. Thus, implicit priors concerning pathology statuses are effectively extracted from the source images and injected into the virtual staining framework.

At the output end, we propose enhancing pathological fidelity aided by PathCLIP. Concretely, we apply PathCLIP's image encoder to both the generated image $\hat{y}$ and the ground-truth target-domain image y. Then, we compute the cosine similarity between the resulting image embeddings and penalize low similarity with the loss

$$\mathcal{L}_{\text{Expert}} = 1 - f_P(\hat{y}) \cdot f_P(y) / \|f_P(\hat{y})\| \, \|f_P(y)\|, \tag{1}$$

where f_P denotes PathCLIP's image encoder and $\|\cdot\|$ denotes L2 norm.

3.3 Enhance Structural and Semantic Consistency with Medical Prior

SD-Turbo (Sauer et al., 2024) and pix2pix-Turbo (Parmar et al., 2024) have applied pixel-wise difference losses for effective training of structure-preserving few-step diffusion models. Pix2pix-Turbo additionally incorporates the learned perceptual image patch similarity (LPIPS) loss (Zhang et al., 2018) to enhance perceptual fidelity. Following the successful practice, we impose a mean squared error (MSE) loss in the latent space and an LPIPS loss in the pixel space:

$$\mathcal{L}_{\text{Struct}} = \omega_M \cdot \mathcal{L}_{\text{MSE}}(\hat{z}, z) + \omega_L \cdot \mathcal{L}_{\text{LPIPS}}(\hat{y}, y), \tag{2}$$

where $\hat{z}$ is the latent predicted by the denoising network, $z = f_V(y)$ is y's latent representation encoded by the VAE encoder f_V, and ω_L and ω_M denote the weights for the two losses. In addition, based on the *medical prior that the H channel highlights nuclei in H&E and IHC images*, we extract the H channels from the generated and ground truth target-domain images via color deconvolution (Ruifrok, 2001), and impose extra consistency losses to enhance structure and pathology fidelity:

$$\mathcal{L}^H_{\text{Struct}} = \omega^H_M \cdot \mathcal{L}^H_{\text{MSE}}(\hat{y}^H, y^H) + \omega^H_L \cdot \mathcal{L}^H_{\text{LPIPS}}(\hat{y}^H, y^H), \tag{3}$$

where $\hat{y}^H$ and y^H denote H channels of $\hat{y}$ and y, respectively, ω^H_M and ω^H_L are weights. This targeted supervision, driven by medical prior, explicitly enhances nuclear structures in the generated images.

3.4 Adversarial Training

Adversarial learning has demonstrated prominent effectiveness in reducing the sampling steps of diffusion models (Parmar et al., 2024; Sauer et al., 2024; Xu et al., 2024). To adapt the text-to-image

generator SD-Turbo (Sauer et al., 2024) for single-step pathology I2I translation, we also incorporate a vision-aided GAN (Kumari et al., 2022) with a frozen DINOv2 (Oquab et al., 2023) backbone, where the discriminator f_D aims to maximize

$$\mathcal{L}_{\text{GAN}}^{D} = \log f_D(y) + \log[1 - f_D(\hat{y})], \tag{4}$$

while the single-step generator f_G (comprising the VAE and denoising network) is trained to minimize

$$\mathcal{L}_{\text{GAN}} = -\log f_D(\hat{y}) = -\log f_D\big(f_G(x)\big). \tag{5}$$

Thus, the generator learns to produce pathology images that match the target-domain distribution, i.e., indistinguishable from real target-domain images by the discriminator. Collectively, the total loss for end-to-end training of the generator is:

$$\mathcal{L} = \omega_E \cdot \mathcal{L}_{\text{Expert}} + \mathcal{L}_{\text{Struct}} + \mathcal{L}_{\text{Struct}}^{H} + \omega_G \cdot \mathcal{L}_{\text{GAN}}, \tag{6}$$

where ω_E and ω_G are constant weights.

3.5 Adapt for High-Resolution Single-Step Staining

Given the ultra-large scale of WSIs, high-resolution generation is desirable to reduce the number of patches required to represent a complete WSI. However, diffusion methods, while capable of generating high-quality images, may suffer from prohibitive computational costs and convergence difficulties when scaled to high-resolution training. Meanwhile, although models trained at a resolution of 512×512 pixels can be directly used to infer 1024×1024-pixel images, performance may be suboptimal due to a lack of adaptation to higher-resolution representations.

We observe that models trained on 512×512-pixel images already acquire core virtual staining abilities, i.e., correctly transferring pathology and structure in general. Yet, their generalization to higher resolutions is limited due to incompatible latent spaces. Hence, such models only need lightweight adaptation for high resolutions. Accordingly, we propose a progressive strategy to train a 1024-resolution virtual staining model efficiently. In the first stage, we train the model on 512×512-pixel images, where the lower computational demand and task complexity allow for rapid convergence. This stage fully fine-tunes the denoising network to enable effective learning of cellular structures and pathological semantics, which are fundamental to the virtual staining task[1]. Then, in the second stage, we introduce LoRA (with rank=8, following (Parmar et al., 2024)) into the denoising network and fine-tune only the LoRA weights using 1024×1024-pixel images. The rationale is to adapt the model for high-resolution quickly with only a small number of parameters, but not interfering with the core virtual staining capabilities acquired in the first stage. Empirical results in Table 3 demonstrate that with the same training epochs, our strategy consistently achieves superior virtual staining quality.

4 Experiments

Datasets and Evaluation Metrics. We evaluate MagicStain on eight datasets for comprehensive performance assessment across various virtual staining tasks. RegH2I (Peng et al., 2024) contains 2,592 pairs of registered images for H&E-to-IHC staining, where the IHC images are labeled with the HER2 biomarker (H&E2IHC-HER2). (Hu et al., 2024) provided 5,098 aligned image pairs for FFPE-to-H&E staining (FFPE2H&E). H&E2IHC-CK, a dataset we collected, comprises 2,557 aligned image pairs for H&E-to-IHC staining, utilizing high molecular weight cytokeratin (CK) as the IHC biomarker. We follow the official train/validation/test splits for FFPE2H&E and H&E2IHC-HER2. For H&E2IHC-CK, we randomly allocate 241 pairs as the test set and use the remaining 2,316 pairs for training. We also evaluate MagicStain on BCI (Liu et al., 2022) and four MIST (Li et al., 2023a) datasets (please refer to Appendix B.8 and Table 11 for details). The H&E2IHC-CK dataset was collected from a prostate cancer cohort, and the other datasets were from breast cancer patient cohorts. Therefore, MagicStain is evaluated with different organs and cancer types. All evaluations are conducted on cropped patches of 1024×1024 pixels in accordance with (Hu et al., 2024; Peng et al., 2024).

[1]Our empirical results showed that fine-tuning the VAE led to similar performance (cf. Appendix B.5 for details). Therefore, we keep the pretrained VAE frozen to reduce training cost.

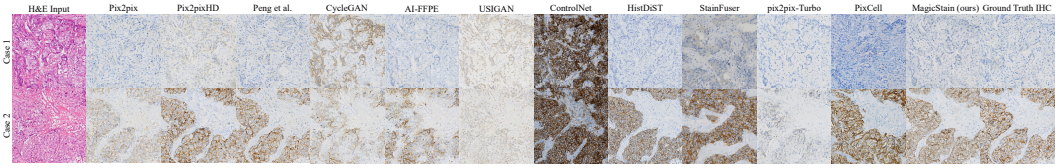

Figure 3: Example virtual staining results of the H&E2IHC-HER2 dataset (best viewed when zoomed in). In the IHC images, brown regions indicate cancerous lesions, with darker brown coloration indicating more pronounced malignancy and a more advanced disease stage.

Following existing practice (Hu et al., 2024; Peng et al., 2024), we report the structural similarity index measure (SSIM) and peak signal-to-noise ratio (PSNR) to reflect the quality of virtually stained images. Typically, higher SSIM values indicate a better structural match (e.g., alignment of nuclear locations and sizes) between the virtually stained and ground-truth images. However, due to intrinsic structural discrepancies in paired pathology images stained with different dyes, exact pixel-wise structural correspondence is generally infeasible, even after registration (cf. Appendix A). This phenomenon is more pronounced in the FFPE2H&E dataset. Therefore, we also report the Fréchet inception distance (FID) (Heusel et al., 2017) and kernel inception distance (KID) (Bińkowski et al., 2018), which are better aligned with human perceptual quality, for a more comprehensive evaluation. Notably, lower FID and KID values indicate a smaller discrepancy between the distributions of generated images and those of the target domains, implying relatively higher quality and fidelity of the virtually stained images compared to real target-domain ones.

Implementation. All experiments are conducted using Ubuntu 20.04, Python 3.10.16, and PyTorch 2.4.0 (Paszke et al., 2019) on a GPU with 80 GB memory. Following pix2pix-Turbo (Parmar et al., 2024), we adopt pretrained SD-Turbo (Sauer et al., 2024) as the backbone of our MagicStain. We use the AdamW optimizer (Loshchilov & Hutter, 2017) with a learning rate of 5×10^{-6}. We train the model with batch sizes of two and one for the resolutions of 512×512 and 1024×1024 pixels, respectively. We first train the model for 15 epochs using images of 512×512 pixels and then fine-tune it for five additional epochs with images of 1024×1024 pixels, where the former is obtained by evenly dividing the latter into four patches. The loss weights are empirically set (cf. Appendix B.6) as follows: $\omega_M = 1$; $\omega_L = 4$; $\omega_L^H = 4$; $\omega_M^H = 0.5$; $\omega_E = 0.2$; and $\omega_G = 0.5$. Our code and trained models will be publicly available.

Comparison with State-of-the-Art (SOTA). We compare MagicStain with classical GAN-based image translation methods: CycleGAN (Zhu et al., 2017), pix2pix (Isola et al., 2017), and pix2pixHD (Wang et al., 2018); SOTA diffusion models for image translation: ControlNet (Zhang et al., 2023) and pix2pix-Turbo (Parmar et al., 2024); and specialized SOTA approaches to virtual staining[2] , including both GAN-based: (Peng et al., 2024), (Hu et al., 2024), AI-FFPE (Ozyoruk et al., 2021), and USIGAN (Peng et al., 2026), and diffusion-based: StainFuser (Jewsbury et al., 2024), HistDiST (Großkopf et al., 2025), and PixCell (Yellapragada et al., 2025). As ControlNet, StainFuser, HistDiST, and PixCell are multi-step diffusion models, we adopt their default inference steps (50, 20, 200, and 20, respectively). Comparisons with (Peng et al., 2024) and (Hu et al., 2024) are exclusively on H&E2IHC (including -HER2 and -CK) and FFPE2H&E, respectively, as each is tailored to its corresponding task.

Quantitative analysis. As Table 1 shows, MagicStain achieves the best performance across all metrics on the H&E2IHC-HER2 and H&E2IHC-CK datasets, indicating superior perceptual and structural quality of its virtually stained images. Quantitative results on FFPE2H&E are deferred to Appendix B.2 (Table 4), where MagicStain remains highly competitive (among the top three) in PSNR and SSIM and again achieves the best FID and KID, demonstrating its superior capability for high-fidelity, high-quality generation. We also benchmark the inference time of the compared diffusion models for generating one sample: ControlNet 18.49 seconds, StainFuser 3.69 seconds, HistDiST 12.27 seconds, pix2pix-Turbo/VIMs 0.439 seconds, PixCell 4.46 seconds, and our MagicStain 0.430 seconds. This comparison highlights the advantage of single-step diffusion models in terms of inference efficiency (see Appendix B.10 for full computational cost).

[2]In our context, the virtual staining model VIMs (Dubey et al., 2024) embodies a straightforward application of pix2pix-Turbo. Therefore, we list them as a single entry (pix2pix-Turbo/VIMs) in our comparison.

Table 1: Evaluation of various methods on the H&E2IHC-HER2, H&E2IHC-Ext, and H&E2IHC-CK datasets. For H&E2IHC-Ext, the models trained on H&E2IHC-HER2 are directly evaluated without any tuning. *: $p < 0.05$ by the Wilcoxon signed-rank test for pairwise comparison with our method.

	H&E2IHC-HER2				H&E2IHC-Ext				H&E2IHC-CK			
Method	FID↓	KID↓	PSNR↑	SSIM↑	FID↓	KID↓	PSNR↑	SSIM↑	FID↓	KID↓	PSNR↑	SSIM↑
Pix2pix	47.77	0.0237	18.04*±3.693	0.401*±0.126	147.0	0.1065	16.63±3.442	0.325*±0.126	51.49	0.0026	19.12*±2.695	0.489*±0.103
Pix2pixHD	41.77	0.0123	18.06*±3.733	0.386*±0.128	104.9	0.0597	16.36±3.390	0.341±0.128	45.44	0.0014	19.10*±2.921	0.482*±0.111
(Peng et al., 2024)	33.92	0.0058	18.02*±3.706	0.385*±0.125	104.7	0.0796	15.70*±2.234	0.345±0.130	50.01	0.0028	18.94*±2.668	0.476*±0.103
CycleGAN	40.91	0.0062	17.01*±3.524	0.365*±0.119	149.6	0.1168	**16.91±3.484**	0.341±0.117	190.3	0.1427	15.73*±3.427	0.443*±0.122
AI-FFPE	37.15	0.0047	17.46*±3.764	0.385*±0.122	92.89	0.0365	16.66±3.235	0.323*±0.122	57.86	0.0027	17.32*±3.297	0.440*±0.120
USIGAN	55.33	0.0249	18.34*±3.766	0.387*±0.117	128.1	0.0670	16.72±3.248	0.332*±0.120	80.71	0.0243	17.06*±3.099	0.384*±0.123
ControlNet	209.7	0.2560	7.368*±0.591	0.127*±0.021	235.9	0.2741	7.245*±0.491	0.107*±0.021	110.0	0.0498	12.20*±1.052	0.313*±0.076
HistDiST	70.99	0.0389	16.82*±3.789	0.390*±0.131	110.2	0.0526	16.00±3.456	0.336*±0.133	62.68	0.0094	17.61*±3.004	0.416*±0.116
StainFuser	123.7	0.1057	14.35*±2.486	0.359*±0.130	182.7	0.1438	13.90*±2.339	0.305*±0.130	104.4	0.0400	12.10*±2.595	0.308*±0.147
pix2pix-Turbo /VIMs	225.2	0.2909	17.91*±3.316	0.344*±0.114	289.2	0.3261	16.61±3.413	0.328*±0.116	56.36	0.0050	18.40*±2.997	0.489*±0.114
PixCell	47.77	0.0138	16.07*±3.595	0.346*±0.127	102.1	0.0387	15.38*±2.947	0.315*±0.119	90.77	0.0271	14.45*±3.424	0.337*±0.154
MagicStain (ours)	**33.51**	**0.0033**	**18.58±3.668**	**0.411±0.125**	**92.24**	**0.0271**	16.17±3.229	**0.348±0.135**	**44.70**	**0.0011**	**19.49±3.017**	**0.505±0.110**

Qualitative analysis. Additionally, we present qualitative results for all three datasets, providing insightful analysis. Figure 3 shows the results on H&E2IHC-HER2 (the results on H&E2IHC-CK and FFPE2H&E are provided in Appendix B.1). As we can see, MagicStain outperforms other methods in structural accuracy by exhibiting more precise cellular structures and clearer cell morphology. This is notably demonstrated in Case 1, where blue dots denote cell nuclei in the IHC-stained image: the distributions and sizes of the nuclei in MagicStain are most consistent with those in the ground truth. Furthermore, MagicStain displays more faithful pathological characteristics. This is evident in Case 2, where brown regions in the IHC images reflect cancerous lesions, with darker brown indicating more pronounced lesions and more advanced disease stages. MagicStain closely matches the ground truth in both lesion location and cancer severity. In contrast, compared methods either underestimate (e.g., pix2pix-Turbo) or overestimate (e.g., ControlNet) the presence/severity of cancer, which is worsened by imprecise lesion expressions. The unsatisfactory pathology statuses of the images generated by ControlNet and pix2pix-Turbo result in poor performance on the downstream task (Table 2). In conclusion, the promising structural and pathological fidelity of MagicStain implies the effectiveness of incorporating domain PFM and a medical prior.

External Validation. To further evaluate the generalizability of our method, we conduct additional validation on the external test set from (Peng et al., 2024), with 285 H&E-IHC image pairs stained with SP3 (D'Alfonso et al., 2013) or CB11 (Purdie et al., 2010) antibodies (H&E2IHC-Ext). Notably, we directly apply the models trained on H&E2IHC-HER2 in the main experiments without further fine-tuning. As shown in Table 1, MagicStain again achieves the best results in FID, KID, and SSIM on H&E2IHC-Ext, while remaining comparable with other methods in PSNR. These results align well with the main experiments, further validating the strong generalizability of MagicStain in virtual staining tasks.

Evaluation on Downstream Task. To examine the pathological fidelity of MagicStain and its practical usability in downstream clinical tasks, we conduct a membrane staining intensity classification experiment using the H&E2IHC-HER2 dataset, following the HER2 scoring guideline (Ahn et al., 2020) and the protocol in (Peng et al., 2024). This task involves categorizing membrane staining intensity into four levels. For each virtual staining method, we use 600 patches of virtual IHC images re-stained from the H&E images in the test set of H&E2IHC-HER2, to train a ResNet-50 classifier (He et al., 2016). The classifier is subsequently evaluated on 285 real IHC images, also from the H&E2IHC-HER2 test set but not paired with the 600 H&E images, to compute accuracy and F1 score. For reference, we have also trained a classifier using H&E images and another using real IHC images. As shown in Table 2, the classifier trained on IHC images generated by MagicStain outperforms those trained on images by all compared methods with large margins. In addition, it is comparable to the upper-bound classifier trained on real IHC images. Meanwhile, the classifier trained on H&E images performs poorly, as expected, due to its inability to convey membrane staining intensity. We also evaluate these classifiers in an out-of-domain setting in Appendix B.13, where MagicStain again beats all competitors. These findings underscore the clinical relevance and potential of accurate virtual staining from H&E to IHC, particularly in facilitating reliable assessment of HER2 expression.

Ablation Studies. We ablate various designs of MagicStain on the HER2-HE2IHC dataset to validate their rationality, and collectively present the results in Table 3. Concerning "pathology expert prior",

Table 2: Evaluation of a downstream IHC classification task.

Method	ACC↑	F1↑
H&E	0.4842	0.3500
Real IHC	**0.8456**	0.8160
Pix2pix	0.6561	0.6183
Pix2pixHD	0.6903	0.6339
(Peng et al., 2024)	0.7929	0.7742
CycleGAN	0.5930	0.5325
AI-FFPE	0.8035	0.7410
USIGAN	0.6526	0.5981
ControlNet	0.4140	0.3836
HistDiST	0.8000	0.7364
StainFuser	0.6561	0.5741
pix2pix-Turbo/VIMs	0.3754	0.3239
PixCell	0.7439	0.7257
w/o prior	0.7895	0.7743
w/o $\mathcal{L}_{\text{Expert}}$	0.8105	0.7974
MagicStain (ours)	0.8421	**0.8171**

Table 3: Ablation study on the H&E2IHC-HER2 dataset. *: $p < 0.05$ by the Wilcoxon signed-rank test for pairwise comparison with our method.

Method	FID↓	KID↓	PSNR↑	SSIM↑
w/o prior	37.48	0.0094	18.38*±3.736	0.411±0.130
w/o $\mathcal{L}_{\text{Expert}}$	35.60	0.0066	18.24*±3.672	0.403*±0.129
w/o $\mathcal{L}_{\text{MSE}}$	34.40	0.0046	18.38*±3.700	0.405*±0.125
w/o $\mathcal{L}_{\text{LPIPS}}$	34.98	0.0049	18.20*±3.580	0.387*±0.124
w/o $\mathcal{L}^H_{\text{MSE}}$	35.85	0.0070	18.01*±3.489	0.396*±0.127
w/o $\mathcal{L}^H_{\text{LPIPS}}$	34.47	0.0063	17.98*±3.506	0.390*±0.125
w/ $\mathcal{L}^I_{\text{MSE}}$	34.40	0.0042	18.51*±3.580	0.409*±0.123
w/o 1024	43.88	0.0135	18.40*±3.885	0.409*±0.133
w/o 512	35.01	0.0053	18.41*±3.587	0.408*±0.128
w/o LoRA	38.15	0.0069	18.50*±3.707	**0.417±0.131**
MagicStain (ours)	**33.51**	**0.0033**	**18.58±3.668**	0.411±0.125

we propose injecting pathology priors extracted by the specialist PFM into the denoising network via multi-head cross-attention. To evaluate this prior's efficacy, we replace it with an empty embedding (denoted by "w/o prior"). We also ablate the PFM-based expert similarity loss $\mathcal{L}_{\text{Expert}}$ (Eqn. (1)) by setting $\omega_E = 0$ (w/o $\mathcal{L}_{\text{Expert}}$). Table 3 shows that removing either the pathology prior or the associated loss $\mathcal{L}_{\text{Expert}}$ degrades FID and KID, while Table 2 shows that their removal also harms downstream task performance due to inaccurate pathological information transfer between the source and target domains, confirming the value of the expertise provided by the specialist PFM.

To investigate the contributions of the structure and pathology consistency losses, we ablate the MSE and LPIPS losses imposed on the entire images ("w/o $\mathcal{L}_{\text{MSE}}$" and "w/o $\mathcal{L}_{\text{LPIPS}}$"), and the corresponding losses on the H-channel ("w/o $\mathcal{L}^H_{\text{MSE}}$" and "w/o $\mathcal{L}^H_{\text{LPIPS}}$"), by setting their weights to 0. Results show that removing any of them consistently degrades all evaluation metrics. In particular, removing MSE and LPIPS losses on the H-channel, which highlights nuclei in the target-domain images, leads to more pronounced drops in PSNR and SSIM. Meanwhile, removing the corresponding losses on the entire image also degrades these metrics. Thus, both types of structural and pathological supervision are necessary for the optimal performance of MagicStain.

For $\mathcal{L}_{\text{MSE}}$, we further experiment with computing MSE in the image space ("w/ $\mathcal{L}^I_{\text{MSE}}$"), instead of the latent space. We find that latent-space supervision yields slightly better results, likely because feature values in the latent space span a broader range, thus providing stronger gradient signals.

For adaptation to high-resolution single-step staining, we propose a two-stage progressive training strategy: first, fine-tuning all parameters on 512×512-pixel images for 15 epochs, followed by LoRA fine-tuning on 1024×1024-pixel images for an additional 5 epochs. We compare this strategy with two alternatives: training only with images of 512×512 or 1024×1024 pixels (w/o 1024 or 512), for 20 epochs, with all parameters updated. Results show that training at a single resolution consistently degrades performance: training only at 512 resolution leads to generalization issues when applied to 1024 resolution, while training only at 1024 resolution is difficult to converge. These results validate our design for high-resolution single-step staining.

Lastly, to determine the necessity of LoRA, we also ablate it by updating all parameters in the second (high-resolution adaptation) training stage (w/o LoRA). As we can see, while PSNR and SSIM remain comparable, both FID and KID degrade substantially. This suggests that, compared with LoRA, fine-tuning all parameters during high-resolution adaptation may cause the model to lose its fundamental virtual staining capability acquired in the first stage.

5 Conclusion

This paper presented MagicStain, a high-fidelity pathology image virtual staining model via single-step diffusion guided by a specialized PFM. Thorough experiments validated its promising performance and effective design. We plan to extend MagicStain to more virtual staining tasks. We discuss limitations, failure cases, and future directions in Appendices E and B.16.

AI USE STATEMENT

No generative AI tools were used for conceptualization, ideation, experimental design, data analysis, or manuscript writing. All theoretical development, experimental design, data analysis, and manuscript text were produced solely by the authors without substantive assistance from AI-based tools. Minor editorial proofreading and grammar checks were limited to standard non-generative tools (e.g., spell-checkers). All scientific content and claims remain the sole responsibility of the authors.

ETHICS STATEMENT

This work presents MagicStain, an application-oriented framework for efficient and high-fidelity virtual staining of pathology images. The broader impact of this work lies in its potential to advance both computational pathology research and AI-driven clinical workflows by offering a practical, scalable, and computationally efficient alternative to conventional chemical staining. MagicStain enhances pathological consistency and visual fidelity, both of which are critical for real-world adoption in clinical pathology. As with any AI-driven medical tool, however, ethical and safety considerations—including potential domain bias, the risk of over-reliance on automated image synthesis, and the necessity of expert oversight—must be carefully addressed prior to any clinical deployment. In particular, MagicStain is currently a research step toward efficient and high-resolution virtual staining rather than a clinically validated diagnostic tool, and without proper qualification and approval from professional regulatory organizations, it should not fully replace conventional chemical staining in clinical practice. This work involves no clinical trials, no interventions with human subjects, and no new dataset releases beyond those already authorized by their respective sources. We foresee no potential for misuse, and the work raises no concerns related to privacy, security, discrimination, fairness, or research integrity. There are no conflicts of interest or external sponsorships associated with this research. Future research should address generalizability across diverse institutions, staining protocols, and clinical settings to ensure reliable and equitable deployment in healthcare. We refer the reader to Appendix D for a broader discussion.

REPRODUCIBILITY STATEMENT

We are committed to ensuring the reproducibility of MagicStain. Upon acceptance, we will publicly release the model weights, the training code, and the inference code under a permissive open-source license, enabling the community to fully reproduce all quantitative and qualitative results reported in this paper. Detailed descriptions of the datasets, evaluation metrics, and implementation details are provided in Section 4 and Section 4 of the main paper. The network architecture, pathology-prior design, loss functions, and the two-stage progressive training strategy are described in Section 3.2 and the subsequent method subsections. Additional experimental settings, ablation studies on loss weights, and dataset-specific details are provided in Appendices A, B.5, and B.6.

REFERENCES

Soomin Ahn, Ji Won Woo, Kyoungyul Lee, and So Yeon Park. HER2 status in breast cancer: changes in guidelines and complicating factors for interpretation. *Journal of Pathology and Translational Medicine*, 54(1):34–44, 2020.

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.

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.

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. MultiDiffusion: fusing diffusion paths for controlled image generation. In *Proceedings of the International Conference on Machine Learning*. JMLR, 2023.

Mahesh Bhosale, Abdul Wasi, Yuanhao Zhai, Yunjie Tian, Samuel Border, Nan Xi, Pinaki Sarder, Junsong Yuan, David Doermann, and Xuan Gong. PathDiff: Histopathology image synthesis with unpaired text and mask conditions. *arXiv preprint arXiv:2506.23440*, 2025.

Chang Bian, Beth Phillips, Tim Cootes, and Martin Fergie. HEMTI: H&E to multiplex-immunohistochemistry image translation with dual-branch pix2pix generator. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 184–197. Springer, 2024.

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. *arXiv preprint arXiv:1801.01401*, 2018.

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 *International Conference on Medical Image Computing and Computer Assisted Intervention*, pp. 384–394. Springer, 2024a.

Fuqiang Chen, Ranran Zhang, Wanming Hu, Deboch Eyob Abera, Yue Peng, Boyun Zheng, Yiwen Sun, Jing Cai, and Wenjian Qin. PGVMS: A prompt-guided unified framework for virtual multiplex IHC staining with pathological semantic learning. *arXiv preprint arXiv:2602.23292*, 2026.

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*, 2024b.

Zizhi Chen, Xinyu Zhang, Minghao Han, Yizhou Liu, Ziyun Qian, Weifeng Zhang, Xukun Zhang, Jingwei Wei, and Lihua Zhang. VLM-based prompts as the optimal assistant for unpaired histopathology virtual staining. *arXiv preprint arXiv:2504.15545*, 2025.

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.

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.

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*, pp. 447–456. Springer, 2023.

Shikha Dubey, Yosep Chong, Beatrice Knudsen, and Shireen Y Elhabian. VIMs: virtual immunohistochemistry multiplex staining via text-to-stain diffusion trained on uniplex stains. In *International Workshop on Machine Learning in Medical Imaging*, pp. 143–155. Springer, 2024.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014.

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

Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Yefei He, Jing Liu, Weijia Wu, Hong Zhou, and Bohan Zhuang. Efficientdm: Efficient quantization-aware fine-tuning of low-bit diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024a.

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 Transactions on Medical Imaging*, 43 (10):3634–3647, 2024b.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in Neural Information Processing Systems*, 30:6629–6640, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

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. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4382–4391. IEEE, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations*, 2022.

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 *International Conference on Medical Image Computing and Computer Assisted Intervention*, pp. 67–76. Springer, 2024.

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.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1125–1134, 2017.

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.

Yifan Jia, Xingda Yu, Zhengyang Ji, Songning Lai, and Yutao Yue. Adaptive H&E-IHC information fusion staining framework based on feature extra. *arXiv preprint arXiv:2502.20156*, 2025.

Mingzhou Jiang, Jiaying Zhou, Nan Zeng, Mickael Li, Qijie Tang, Chao He, Huazhu Fu, and Honghui He. Topology-aware pathological consistency matching for weakly-paired IHC virtual staining. *arXiv preprint arXiv:2601.02806*, 2026.

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

Tushar Kataria, Beatrice Knudsen, and Shireen Y Elhabian. Implicitstainer: Data-efficient medical image translation for virtual antibody-based tissue staining using local implicit functions. *arXiv preprint arXiv:2505.09831*, 2025.

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.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes, 2022. URL `https://arxiv.org/abs/1312.6114`.

Pascal Klöckner, José Teixeira, Diana Montezuma, Jaime S Cardoso, Hugo M Horlings, and Sara P Oliveira. GANs vs. diffusion models for virtual staining with the HER2match dataset. *arXiv preprint arXiv:2506.18484*, 2025.

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.

Nupur Kumari, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Ensembling off-the-shelf models for GAN training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10651–10662, 2022.

Junhyeok Lee, Han Jang, Heeseong Eum, Joon Jang, and Kyu Sung Choi. Segmentation-before-staining improves structural fidelity in virtual IHC-to-Multiplex IF translation. *arXiv preprint arXiv:2603.16160*, 2026.

Fangda Li, Zhiqiang Hu, Wen Chen, and Avinash Kak. Adaptive supervised PatchNCE loss for learning H&E-to-IHC stain translation with inconsistent groundtruth image pairs. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 632–641. Springer, 2023a.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11259–11268, 2024a.

Xiuyu Li, Yijiang Liu, Long Lian, Huanrui Yang, Zhen Dong, Daniel Kang, Shanghang Zhang, and Kurt Keutzer. Q-diffusion: Quantizing diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17535–17545, 2023b.

Yanjing Li, Sheng Xu, Xianbin Cao, Xiao Sun, and Baochang Zhang. Q-DM: An efficient low-bit quantized diffusion model. *Advances in Neural Information Processing Systems*, 36:76680–76691, 2023c.

Yuzhu Li, Nir Pillar, Jingxi Li, Tairan Liu, Di Wu, Songyu Sun, Guangdong Ma, Kevin de Haan, Luzhe Huang, Yijie Zhang, et al. Virtual histological staining of unlabeled autopsy tissue. *Nature Communications*, 15(1):1684, 2024b.

Anran Liu, Xiaofei Wang, Jing Cai, and Chao Li. Score-based diffusion model for unpaired virtual histology staining. *arXiv preprint arXiv:2506.23184*, 2025.

Shengjie Liu, Chuang Zhu, Feng Xu, Xinyu Jia, Zhongyue Shi, and Mulan Jin. BCI: Breast cancer immunohistochemical image generation through pyramid pix2pix. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1815–1824, 2022.

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 Transactions on Medical Imaging*, 40(8):1977–1989, 2021.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Zhijie Liu, Lingchao Chen, Haixia Cheng, Jianpeng Ao, Ji Xiong, Xing Liu, Yaxin Chen, Ying Mao, and Minbiao Ji. Virtual formalin-fixed and paraffin-embedded staining of fresh brain tissue via stimulated Raman CycleGAN model. *Science Advances*, 10(13):eadn3426, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

Jiabo Ma, Wenqiang Li, Jinbang Li, Ziyi Liu, Linshan Wu, Fengtao Zhou, Li Liang, Ronald Cheong Kin Chan, Terence TW Wong, and Hao Chen. Generative AI for misalignment-resistant virtual staining to accelerate histopathology workflows. *Nature Communications*, 2026.

Tao Ma, Chao Zhang, Min Lu, and Lin Luo. AGMDT: Virtual staining of renal histology images with adjacency-guided multi-domain transfer. *arXiv preprint arXiv:2309.06421*, 2023.

Marcin Możejko, Dawid Uchal, Krzysztof Gogolewski, Piotr Kupidura, Szymon Łukasik, Jakub Giezgała, Tomasz Nocoń, Kacper Pietrzyk, Robert Pieniuta, Mateusz Sulimowicz, et al. ImmuVis: Hyperconvolutional foundation model for imaging mass cytometry. *arXiv preprint arXiv:2602.04585*, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Kutsev Bengisu Ozyoruk, Sermet Can, Guliz Irem Gokceler, Kayhan Basak, Derya Demir, Gurdeniz Serin, Uguray Payam Hacisalihoglu, Emirhan Kurtuluş, Berkan Darbaz, Ming Y Lu, et al. Deep learning-based frozen section to FFPE translation. *arXiv preprint arXiv:2107.11786*, 2021.

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.

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. *Advances in Neural Information Processing Systems*, 32: 8026–8037, 2019.

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*, 6(9):1077–1093, 2024.

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 *International Conference on Medical Image Computing and Computer Assisted Intervention*, pp. 3–13. Springer, 2024.

Yue Peng, Bing Xiong, Fuqiang Chen, Deboch Eybo Abera, RanRan Zhang, Wanming Hu, Jing Cai, and Wenjian Qin. USIGAN: Unbalanced self-information feature transport for weakly paired image IHC virtual staining. *IEEE Transactions on Image Processing*, 2026.

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.

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 *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Jillur Rahman Saurav, Mohammad Sadegh Nasr, and Jacob M Luber. Cross-domain image synthesis: Generating H&E from multiplex biomarker imaging. *arXiv e-prints*, pp. arXiv–2508, 2025.

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 *IEEE International Conference on Machine Learning and Applications*, pp. 828–834. IEEE, 2018.

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.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Arnout Ruifrok. Quantification of histochemical staining by color deconvolution. *Analytical and quantitative cytology and histology/the International Academy of Cytology [and] American Society of Cytology*, 2001.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pp. 87–103. Springer, 2024.

Jillur Rahman Saurav, Thuong Le Hoai Pham, Pritam Mukherjee, Paul Yi, Brent A Orr, and Jacob M Luber. UNIStainNet: Foundation-model-guided virtual staining of H&E to IHC. *arXiv preprint arXiv:2603.12716*, 2026.

Tibor Sloboda, Lukáš Hudec, and Wanda Benešová. Editable stain transformation of histological images using unpaired GANs. In *International Conference on Image Analysis and Processing*, pp. 27–38. Springer, 2023.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the International Conference on Machine Learning*, pp. 32211–32252. JMLR, 2023.

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.

Kexin Sun, Zhineng Chen, Gongwei Wang, Jun Liu, Xiongjun Ye, and Yu-Gang Jiang. Bi-directional feature fusion generative adversarial network for ultra-high resolution pathological image virtual restaining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3904–3913, 2023.

Yuxuan Sun, Yunlong Zhang, Yixuan Si, Chenglu Zhu, Zhongyi Shui, Kai Zhang, Jingxiong Li, Xingheng Lyu, Tao Lin, and Lin Yang. PathGen-1.6M: 1.6 million pathology image-text pairs generation through multi-agent collaboration. *arXiv preprint arXiv:2407.00203*, 2024a.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5034–5042, 2024b.

Yudong Tian, Xiangyu Zhao, Yuqing Liu, Bofei Yang, and Chongzhao Wu. Fourier transform infrared microspectroscopy-based super-resolution virtual staining of unlabeled tissues by pixel diffusion transformer. *arXiv preprint arXiv:2603.08143*, 2026.

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*, 2024a.

Song Wang, Zhong Zhang, Huan Yan, Ming Xu, and Guanghui Wang. Mix-domain contrastive learning for unpaired H&E-to-IHC stain translation. In *IEEE International Conference on Image Processing*, pp. 2982–2988. IEEE, 2024b.

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8798–8807, 2018.

Tong Wang, Mingkang Wang, Zhongze Wang, Hongkai Wang, Qi Xu, Fengyu Cong, and Hongming Xu. ODA-GAN: Orthogonal decoupling alignment GAN assisted by weakly-supervised learning for virtual immunohistochemistry staining. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 25920–25929, 2025.

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*, pp. 1–10. Springer, 2024.

Bing Xiong, Yue Peng, RanRan Zhang, Fuqiang Chen, Jiaye He, and Wenjian Qin. Unpaired multi-domain histopathology virtual staining using dual path prompted inversion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8780–8787, 2025.

Siyuan Xu, Haofei Song, Yingjiao Deng, Jiansheng Wang, Yan Wang, and Qingli Li. Advancing stain transfer for multi-biomarkers: A human annotation-free method based on auxiliary task supervision. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence, IJCAI 2025*, pp. 2116–2124. International Joint Conferences on Artificial Intelligence, 2025a.

Yanwu Xu, Yang Zhao, Zhisheng Xiao, and Tingbo Hou. UFOGen: You forward once large scale text-to-image generation via diffusion GANs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8196–8206, 2024.

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.

Ziwang Xu, Lanqing Guo, Satoshi Tsutsui, Shuyan Zhang, Alex C Kot, and Bihan Wen. Digital staining with knowledge distillation: A unified framework for unpaired and paired-but-misaligned data. *IEEE Transactions on Medical Imaging*, 2025b.

Xudong Yan, Mingze Yuan, Yao Lu, Ying Zhang, Zifan Chen, Peng Bao, Zehua Li, Bin Dong, Li Yang, Li Zhang, et al. Versatile stain transfer in histopathology using a unified diffusion framework. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5. IEEE, 2025.

Hao Yang, JianYu Wu, Run Fang, Xuelian Zhao, Yuan Ji, Zhiyu Chen, Guibin He, Junceng Guo, Yang Liu, and Xinhua Zeng. Cross-channel perception learning for H&E-to-IHC virtual staining. *arXiv preprint arXiv:2506.07559*, 2025.

Srikar Yellapragada, Alexandros Graikos, Zilinghan Li, Kostas Triaridis, Varun Belagali, Saarthak Kapse, Tarak Nath Nandi, Ravi K Madduri, Prateek Prasanna, Tahsin Kurc, et al. PixCell: A generative foundation model for digital histopathology images. *ArXiv*, pp. arXiv–2506, 2025.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6613–6623, 2024.

Yizhe Yuan, Bingsen Xue, Bangzheng Pu, Chengxiang Wang, and Cheng Jin. PRINTER: deformation-aware adversarial learning for virtual IHC staining with in situ fidelity. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 10210–10219, 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Wei Zhang, Tik Ho Hui, Pui Ying Tse, Fraser Hill, Condon Lau, and Xinyue Li. High-resolution medical image translation via patch alignment-based bidirectional contrastive learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 178–188. Springer, 2024.

Xiaoyu Zheng, Jing Wen, Jiaxin Zhuang, Yao Du, Jing Cong, Limei Guo, Chao He, Lin Luo, and Hao Chen. Diffusion-based virtual staining from polarimetric mueller matrix imaging. *arXiv preprint arXiv:2503.01352*, 2025.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2223–2232, 2017.

Peide Zhu, Linbin Lu, Zhiqin Chen, and Xiong Chen. LGD-Net: Latent-guided dual-stream network for HER2 scoring with task-specific domain knowledge. *arXiv preprint arXiv:2602.17793*, 2026.

Eric Zimmermann, Eugene Vorontsov, Julian Viret, Adam Casson, Michal Zelechowski, George Shaikovski, Neil Tenenholtz, James Hall, Thomas Fuchs, Nicolo Fusi, Siqi Liu, and Kristen Severson. Virchow2: Scaling self-supervised mixed magnification models in pathology. *arXiv preprint arXiv:2408.00738*, 2024.

A More Explanation of Non-Perfect Paired Data

In virtual staining tasks, paired images are typically obtained from consecutive tissue sections or chemical re-staining of the same tissue section. The former yields two distinct tissue sections, and the latter inevitably deforms the tissue section between the two stains. Therefore, it is generally infeasible to achieve perfect registration. While the alignment does not achieve perfect pixel-wise matching, the majority of pixels are well-aligned, making them suitable for learning structural correspondences. Prior to our approach, numerous methods have effectively trained virtual staining models using the same datasets as in this study (Hu et al., 2024; Peng et al., 2024).

B Additional Experiments

B.1 Additional Qualitative Results

We present additional qualitative results for the H&E2IHC-HER2, H&E2IHC-CK, and FFPE2H&E datasets under the same experimental settings as Tables 1 and 4 and Figure 3 in the main text. Figure 4 shows that MagicStain exhibits clearer tissue structures, more precise cellular morphology, and more faithful pathological characteristics than other methods.

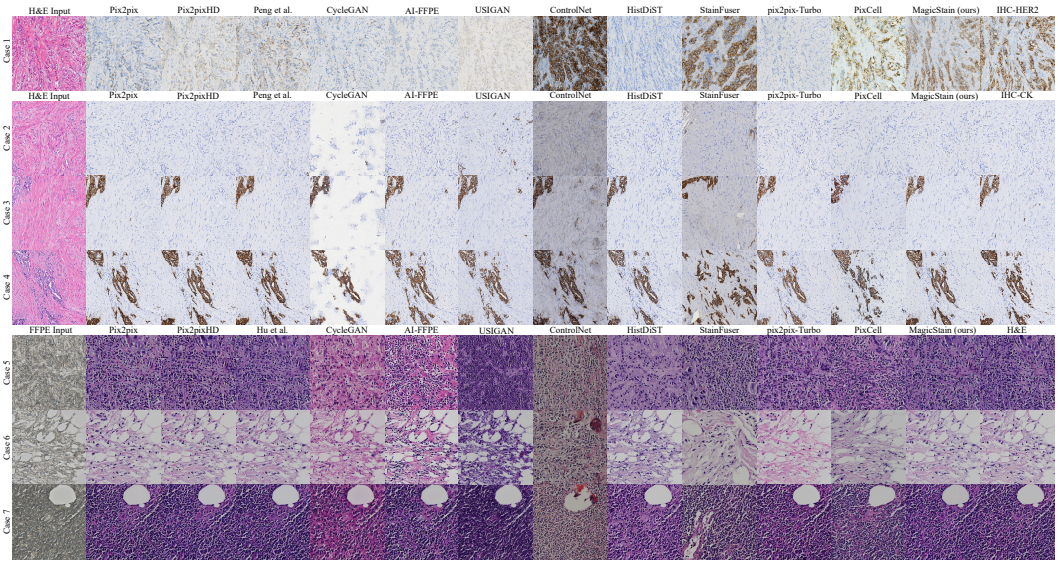

Figure 4: Example virtual staining results by various methods on three tasks: Case 1: H&E2IHC-HER2; Cases 2–4: H&E2IHC-CK; and Cases 5–7: FFPE2H&E (best viewed when zoomed in). In IHC-HER2 images, brown regions indicate cancerous lesions. In IHC-CK images, brown regions indicate epithelial cells. In particular, in the IHC-HER2 image, darker brown coloration signifies higher malignancy and a more advanced disease stage. In H&E images, dark purple dots represent cell nuclei, while light purple or reddish regions correspond to the cytoplasm. Enlarged nuclei may indicate malignant cells.

B.2 Quantitative Results on FFPE2H&E

We report the quantitative evaluation on the FFPE2H&E dataset in Table 4. MagicStain shows slightly lower PSNR and SSIM than the best numbers but remains highly competitive (among the top three) on FFPE2H&E. We speculate that this is due to the relatively weaker pixel- and structure-level correspondence between paired images in FFPE2H&E compared with the other datasets, which affects structural evaluation. Nevertheless, MagicStain achieves the best scores on the perception-oriented metrics—FID and KID—again demonstrating its superior capability for high-fidelity, high-quality generation.

Table 4: Evaluation of various methods on the FFPE2H&E dataset. *: $p < 0.05$ by the Wilcoxon signed-rank test for pairwise comparison with our method.

Method	FID↓	KID↓	PSNR↑	SSIM↑
Pix2pix	17.88	0.0008	18.34*±2.009	0.536*±0.113
Pix2pixHD	15.26	0.0008	19.08±2.046	0.586±0.106
(Hu et al., 2024)	31.56	0.0101	**19.86±2.082**	**0.644±0.098**
CycleGAN	35.36	0.0087	14.47*±2.045	0.363*±0.152
AI-FFPE	25.39	0.0022	13.92*±1.971	0.356*±0.148
USIGAN	90.01	0.0546	12.49*±1.369	0.256*±0.148
ControlNet	156.0	0.1039	11.50*±0.882	0.187*±0.069
HistDiST	60.56	0.0339	15.75*±2.018	0.399*±0.140
StainFuser	44.38	0.0096	12.25*±1.665	0.256*±0.152
pix2pix-Turbo /VIMs	22.33	0.0022	16.93*±1.697	0.505*±0.113
PixCell	31.32	0.0066	12.83*±2.080	0.277*±0.177
MagicStain (ours)	**13.98**	**0.0001**	18.52±2.082	0.544±0.112

Table 5: Performance comparison of using various pathology foundation models (PFMs) as the expert on the H&E2IHC-HER2 dataset.

PFM	FID↓	KID↓	PSNR↑	SSIM↑
UNI2-h	33.84	0.0064	18.50±3.721	0.408±0.126
Virchow2	33.94	0.0050	18.32±3.624	0.402±0.127
PathGenCLIP	33.52	0.0034	18.65±3.750	0.410±0.131
PathCLIP (ours)	33.51	0.0033	18.58±3.668	0.411±0.125

B.3 Choice of Pathology Foundation Model

For the "expert" pathology foundation model (PFM) that provides pathology expertise in MagicStain, we have adopted PathCLIP (Sun et al., 2024b)—a vision-language model (VLM) pretrained via contrastive learning between textual and visual embeddings—in the main text. Here, we further experiment with an additional pathology VLM, PathGenCLIP (Sun et al., 2024a), and two DINO-based self-supervised PFMs, UNI2-h (Chen et al., 2024b) and Virchow2 (Zimmermann et al., 2024). We replace PathCLIP with PathGenCLIP, UNI2-h, or Virchow2 on the H&E2IHC-HER2 dataset for both image-to-image guidance and pathology-accuracy supervision.

Table 5 presents the performance on the H&E2IHC-HER2 dataset. MagicStain with PathGenCLIP, UNI2-h, and Virchow achieve results comparable to those with PathCLIP and outperforms all other methods in Table 1, indicating that MagicStain can effectively incorporate other PFMs to achieve consistently strong performance. Given PathCLIP's overall superior performance compared to other alternatives, we ultimately adopt it to provide the pathology prior.

B.4 Should Both Stages Adopt Low-Rank Adaptation (LoRA)?

In the main text, we have proposed progressive low-to-high-resolution adaptation, where the first stage updates all parameters of the denoising network using 512×512-pixel images, and the second stage implements LoRA using 1024×1024-pixel images. This strategy has demonstrated a major contribution in MagicStain, as revealed by our ablation study in Table 3. Here, we explore the question: can LoRA also be used in the first stage for more computationally efficient training? This means that throughout the low- and high-resolution training, only the LoRA parameters are updated. Experimental results are shown in Table 6. Compared with full-parameter fine-tuning, using LoRA in the first stage leads to substantial performance drops across all evaluation metrics. These results suggest that to endow the baseline text-to-image one-step diffusion model pretrained on natural images with the core virtual staining capabilities, extensive parameter updates rather than LoRA are needed.

Table 6: Comparison between LoRA and full-parameter fine-tuning (ours) in the first training stage (using images of 512×512 pixels) on the H&E2IHC-HER2 Dataset. *: $p < 0.05$ by the Wilcoxon signed-rank test.

First stage	FID↓	KID↓	PSNR↑	SSIM↑
LoRA	77.36	0.0472	17.67*±3.253	0.395*±0.117
Full (ours)	**33.51**	**0.0033**	**18.58**±3.668	**0.411**±0.125

B.5 Influence of Fine-Tuning Variational Autoencoder (VAE)

MagicStain uses the frozen, pretrained VAE from Stable Diffusion (SD) (Rombach et al., 2022) for pathology image encoding and reconstruction. This section empirically studies the impact of fine-tuning the VAE (with LoRA, similar to (Parmar et al., 2024)) on performance using the H&E2IHC-HER2 task. As shown in Table 7, fine-tuning the VAE leads to marginally better FID but worse KID, PSNR, and SSIM. These results suggest that, despite the potential domain gap between natural and pathology images, the pretrained VAE may already suffice for pathology image encoding and reconstruction in MagicStain, and that this gap has minimal impact on performance in this case. We attribute this to the SD VAE's strong encoding/decoding ability acquired from large-scale, diverse training, which likely includes pathology images or similar ones. To reduce training cost, we therefore opt not to train the VAE further.

Table 7: Ablation study results on the H&E2IHC-HER2 dataset evaluating the impact of fine-tuning the VAE.

VAE status	FID↓	KID↓	PSNR↑	SSIM↑
Fine-tuning	**32.33**	0.0048	18.45±3.752	0.410±0.132
Frozen (ours)	33.51	**0.0033**	**18.58**±3.668	**0.411**±0.125

B.6 Ablation Studies on Loss Weights

Our total training loss in Eqn. (6) is dependent on a few weights, so is the model performance. Therefore, we have determined their values to optimize model performance, based on empirical evidence on the H&E2IHC-HER2 dataset. We group the weights into three groups: (i) ω_E and ω_G for the PFM-based expert similarity and GAN losses, (ii) ω_M and ω_L for the original structural and semantic consistency losses in Eqn. (2), and (iii) ω_M^H and ω_L^H for the H-channel enhanced structural and semantic consistency losses in Eqn. (3), and vary either weight in a group while fixing all other ones. The rough candidate ranges for the weights are determined from corresponding or similar weight values in (Parmar et al., 2024). Tables 8, 9, and 10 present the experimental results. As we can see, while PSNR and SSIM show no significant variation across different weight values, FID and KID—which are better aligned with human perceptual quality and reflect global image fidelity—exhibit noticeable differences. Based on the general principle of choosing the weight value that consistently ranks among the top two across all evaluated metrics, we set $\omega_E = 0.2$; $\omega_G = 0.5$; $\omega_M = 1$; $\omega_L = 4$; $\omega_M^H = 0.5$; $\omega_L^H = 4$, which represent a balance of various losses.

B.7 Dataset Organization

Table 11 summarizes the datasets used in this paper, including their tasks, train/test splits, split protocols, and evaluation resolutions.

B.8 Additional Experiments on Public Datasets

We present additional quantitative results on widely used public datasets, including BCI (Liu et al., 2022) and four MIST (Li et al., 2023a) datasets (-HER2, -ER, -Ki67, and -PR), under the same experimental settings as Tables 1 in the main text. **BCI** Liu et al. (2022) includes 3,896 training and 977 testing pairs labeled with the HER2 biomarker. **MIST-HER2** Li et al. (2023a) contains 4,642 training and 1,000 testing pairs, also labeled with HER2. **MIST-ER** Li et al. (2023a) is labeled with

Table 8: The impact of varying ω_M^H and ω_L^H on performance on the H&E2IHC-HER2 dataset. *: $p < 0.05$ by the Wilcoxon signed-rank test for pairwise comparison with $\omega_M^H = 0.5$ (top) and $\omega_L^H = 4$ (bottom).

Weight	FID↓	KID↓	PSNR↑	SSIM↑
$\omega_M^H = 0$	35.85	0.0070	18.01*±3.489	0.396*±0.127
$\omega_M^H = 0.25$	35.34	0.0066	18.36*±3.570	0.409*±0.125
$\omega_M^H = 0.5$	**33.51**	**0.0033**	**18.58**±3.668	**0.411**±0.125
$\omega_M^H = 1$	35.09	0.0046	18.38*±3.582	0.408*±0.125
$\omega_M^H = 2$	34.77	0.0063	18.47*±3.554	0.409*±0.124
$\omega_L^H = 0$	34.47	0.0063	17.98*±3.506	0.390*±0.125
$\omega_L^H = 1$	35.02	0.0045	18.26*±3.611	0.403*±0.127
$\omega_L^H = 2$	34.70	0.0051	18.31*±3.637	0.405*±0.127
$\omega_L^H = 4$	**33.51**	**0.0033**	**18.58**±3.668	0.411±0.125
$\omega_L^H = 8$	35.70	0.0043	18.49*±3.717	**0.415**±0.127

Table 9: The impact of varying ω_E and ω_G on performance on the H&E2IHC-HER2 dataset. *: $p < 0.05$ by the Wilcoxon signed-rank test for pairwise comparison with $\omega_E = 0.2$ (top) and $\omega_G = 0.5$ (bottom).

Weight	FID↓	KID↓	PSNR↑	SSIM↑
$\omega_E = 0$	35.60	0.0066	18.24*±3.672	0.403*±0.129
$\omega_E = 0.1$	34.35	0.0073	18.39*±3.588	0.409*±0.125
$\omega_E = 0.2$	**33.51**	**0.0033**	**18.58**±3.668	**0.411**±0.125
$\omega_E = 0.4$	34.25	0.0072	18.34*±3.568	0.407*±0.125
$\omega_E = 0.8$	34.84	0.0059	18.37*±3.581	0.406*±0.125
$\omega_G = 0.25$	35.43	0.0061	18.57±3.670	**0.412**±0.129
$\omega_G = 0.5$	**33.51**	**0.0033**	**18.58**±3.668	0.411±0.125
$\omega_G = 1.0$	36.31	0.0071	18.17*±3.701	0.403*±0.127
$\omega_G = 2.0$	36.40	0.0056	18.06*±3.678	0.391*±0.125

the ER biomarker and includes 4,153 training and 1,000 testing pairs. **MIST-Ki67** Li et al. (2023a) is labeled with the Ki67 biomarker and provides 4,361 training and 1,000 testing pairs. **MIST-PR** Li et al. (2023a) includes 4,139 training and 1,000 testing pairs labeled with the PR biomarker. As shown in Tables 12, 13, and 14, MagicStain achieves the best performance across most metrics (only slightly lower PSNR compared to Pix2pix on MIST-PR and SSIM compared to StainFuser on MIST-Ki67) on the five public datasets, demonstrating superior perceptual and structural quality in its virtually stained images as well as excellent generalization across datasets.

B.9 Analysis of Training Convergence

To verify that MagicStain converges within 15-20 epochs, we conduct an experiment to train it for 200 epochs (150 for stage 1 and 50 for stage 2), following the same ratio as Section 4 in the main text (15 and 5 epochs). As shown in Tables 15, the 200-epoch model yields results similar to those obtained with 20 epochs, indicating that MagicStain has converged at 20 epochs. This rapid convergence is attributed to MagicStain's domain-specific modules, medically motivated loss functions, and the progressive training strategy designed for high-resolution adaptation. These designs enable the model to achieve high-fidelity, high-resolution virtual staining and surpass Pix2pix (Isola et al., 2017) and CycleGAN (Zhu et al., 2017) that require 200 epochs of training, as demonstrated in Sections 4.

B.10 Computational Cost

We report the training and inference computational cost of various methods in Tables 16 and 17, measured on the same single-GPU hardware described in Section 4 on the H&E2IHC-HER2 dataset, using each method's official batch size and resolution.

Table 10: The impact of varying ω_M and ω_L on performance on the H&E2IHC-HER2 dataset. *: $p < 0.05$ by the Wilcoxon signed-rank test for pairwise comparison with $\omega_M = 1$ (top) and $\omega_L = 4$ (bottom).

Weight	FID↓	KID↓	PSNR↑	SSIM↑
$\omega_M = 0$	34.40	<u>0.0046</u>	18.38*±3.700	0.405*±0.125
$\omega_M = 0.25$	**33.24**	0.0055	18.44*±3.775	0.407*±0.125
$\omega_M = 0.5$	34.02	0.0053	18.40*±3.736	<u>0.412</u>±0.129
$\omega_M = 1$	<u>33.51</u>	**0.0033**	**18.58**±3.668	0.411±0.125
$\omega_M = 2$	34.34	0.0070	**18.58**±3.652	**0.416**±0.128
$\omega_L = 0$	34.98	0.0049	18.20*±3.580	0.387*±0.124
$\omega_L = 1$	35.55	0.0047	18.27*±3.631	0.403*±0.125
$\omega_L = 2$	34.80	<u>0.0041</u>	18.34*±3.715	0.404*±0.129
$\omega_L = 4$	**33.51**	**0.0033**	**18.58**±3.668	<u>0.411</u>±0.125
$\omega_L = 8$	<u>33.85</u>	0.0047	<u>18.48</u>*±3.637	**0.413**±0.127

Table 11: Dataset organization.

Dataset	Task / biomarker	Train / test split	Split protocol	Evaluation resolution
H&E2IHC-HER2	H&E→IHC / HER2	official split	official train/validation/test split	1024×1024
FFPE2H&E	FFPE→H&E	official split	official train/validation/test split	1024×1024
H&E2IHC-CK	H&E→IHC / CK	2,316 / 241	random test split	1024×1024
BCI	H&E→IHC / HER2	3,896 / 977	public split	1024×1024
MIST-HER2	H&E→IHC / HER2	4,642 / 1,000	public split	1024×1024
MIST-ER	H&E→IHC / ER	4,153 / 1,000	public split	1024×1024
MIST-Ki67	H&E→IHC / Ki67	4,361 / 1,000	public split	1024×1024
MIST-PR	H&E→IHC / PR	4,139 / 1,000	public split	1024×1024

The two stages of MagicStain are strictly sequential: the 1024-stage LoRA fine-tunes from the 512-stage checkpoint, so the total training cost is the sum of both stages (16.77h + 8.13h = 24.90h), still below ControlNet (117.37h), StainFuser (52.74h), and PixCell (52.29h). At inference, MagicStain runs at 0.430s/image, ∼43× faster than ControlNet (18.495s), ∼29× faster than HistDiST (12.271s), ∼10× faster than PixCell (4.457s), and comparable to pix2pix-Turbo (0.439s), while consuming the lowest VRAM (5.102 GB).

B.11 Impact of DAB Channel Enhancement

An IHC image comprises two main staining components: a Hematoxylin (H) channel that highlights nuclear morphology and a DAB channel that encodes the spatial distribution of the target biomarker. The brown DAB channel in IHC images highlights the location and intensity of target protein expression (e.g., HER2 on the cell membrane and CK in epithelial cells; darker brown indicates stronger expression), which primarily encodes pathological information. In the main text, MagicStain enhances perceptual and reconstruction losses on the H channel to enforce precise nuclear morphology generation, and relies on PathCLIP (Sun et al., 2024b) to supervise pathological accuracy when generating H&E or IHC images. As demonstrated in Tables 2 and 3, these components are effective.

Here, we also explore the impact of enhancing the DAB channel. We conduct experiments on the H&E2IHC-HER2 dataset by enforcing perceptual and reconstruction losses on the DAB channel instead of the H channel (w DAB w/o H) or by using both (w DAB w H). The results of Table 18 show that jointly enhancing the DAB and H channels yields performance similar to not enhancing the DAB channel (slightly better FID and PSNR, and slightly worse KID and SSIM). We conjecture this is because, MagicStain has already incorporated global supervision on the entire image through an MSE loss in the latent space and an LPIPS loss in the pixel space, which also ensures accurate semantics of its DAB channel. However, removing the H channel enhancement (w DAB w/o H) results in clearly inferior performance, indicating that the H channel enhancement is crucial for MagicStain in terms of precise nuclear morphology.

Furthermore, the DAB channel loss can be used only to generate IHC images and cannot be applied to FFPE-to-H&E staining, thereby limiting the method's applicability across multiple stain types. In contrast, the H channel, which highlights nuclei in both H&E and IHC images, can be used for both

Table 12: Evaluation of various methods on the BCI datasets. *: $p < 0.05$ by the Wilcoxon signed-rank test for pairwise comparison with our method.

Method	FID↓	KID↓	PSNR↑	SSIM↑
Pix2pix	91.06	0.0580	21.28*±4.773	0.544*±0.172
Pix2pixHD	88.81	0.0583	19.47*±5.657	0.532*±0.172
(Peng et al., 2024)	63.18	0.0258	19.96*±5.263	0.517*±0.169
CycleGAN	61.09	0.0304	16.46*±5.858	0.511*±0.166
AI-FFPE	51.31	0.0226	15.09*±4.837	0.491*±0.147
USIGAN	96.85	0.0752	16.36*±3.091	0.436*±0.146
ControlNet	236.9	0.2381	9.10*±2.864	0.248*±0.065
HistDiST	56.12	0.0230	20.12*±5.302	0.537*±0.169
StainFuser	56.67	0.0268	17.22*±3.942	0.541*±0.179
pix2pix-Turbo/VIMs	303.1	0.4019	13.64*±4.494	0.443*±0.160
PixCell	42.24	0.0179	19.92*±3.509	0.527*±0.173
MagicStain (ours)	**30.44**	**0.0041**	**22.51**±4.367	**0.556**±0.182

Table 13: Evaluation of various methods on the MIST-HER2 and MIST-ER datasets. *: $p < 0.05$ by the Wilcoxon signed-rank test for pairwise comparison with our method.

	MIST-HER2				MIST-ER			
Method	FID↓	KID↓	PSNR↑	SSIM↑	FID↓	KID↓	PSNR↑	SSIM↑
Pix2pix	75.97	0.0427	15.02*±2.855	0.249*±0.084	70.53	0.0382	15.071±3.209	0.269*±0.093
Pix2pixHD	61.45	0.0236	14.59*±2.967	0.243*±0.085	58.45	0.0213	14.468*±3.192	0.263*±0.095
(Peng et al., 2024)	50.95	0.0148	14.58*±2.971	0.241*±0.087	54.68	0.0204	14.352*±3.164	0.255*±0.093
CycleGAN	67.78	0.0343	13.91*±2.621	0.228*±0.077	46.66	0.0142	13.113*±2.516	0.247*±0.093
AI-FFPE	57.41	0.0217	13.37*±2.433	0.225*±0.076	72.54	0.0413	14.065*±2.738	0.254*±0.080
USIGAN	37.20	0.0049	14.05*±2.641	0.225*±0.082	33.59	0.0037	13.764*±2.564	0.240*±0.083
ControlNet	170.2	0.1244	8.335*±1.394	0.135*±0.024	245.6	0.2156	7.397*±1.209	0.124*±0.020
HistDiST	77.53	0.0408	13.87*±2.964	0.255±0.090	95.75	0.0654	14.481*±3.259	0.267*±0.093
StainFuser	58.36	0.0228	13.42*±2.531	0.255±0.105	62.79	0.0322	12.542*±2.179	0.276*±0.118
pix2pix-Turbo/VIMs	152.5	0.1018	14.04*±2.634	0.214*±0.072	79.73	0.0371	13.458*±2.406	0.209*±0.079
PixCell	50.22	0.0185	14.36*±2.703	0.232*±0.084	43.88	0.0166	14.222*±3.013	0.253*±0.095
MagicStain (ours)	**35.64**	**0.0031**	**15.20**±2.900	**0.257**±0.091	**30.51**	**0.0023**	**15.114**±3.364	**0.280**±0.104

generation tasks. Therefore, MagicStain incorporates the H-channel loss and PathCLIP similarity loss for H&E and IHC image generation.

B.12 Impact of Pretrained Weights

MagicStain adapts pretrained SD-Turbo (Sauer et al., 2024) as its backbone, like many diffusion applications that use pretrained weights on generic natural images as initialization to achieve better performance. In this section, we also evaluate the performance of training without pretrained weights—otherwise under the same setting as Table 1 in the main text. As shown in Table 19, removing pretrained weights yields worse results than using pretrained initialization. Nevertheless, it still achieves highly competitive performance compared with existing SOTA methods in Table 1 (only inferior to (Peng et al., 2024)), owing to our architectural advancements—including explicit incorporation of domain knowledge and high-resolution adaptation.

B.13 Downstream Task Performance on Out-of-Domain Data

In Section 4, we assess the downstream task performance in an "in-domain" setting, where the classifiers are trained on IHC images in the H&E2IHC-HER2 dataset (either real or virtually stained) and then evaluated on real IHC images in the same dataset. In this section, we further evaluate the classifiers in an out-of-domain setting by testing them on the 285 IHC images in H&E2IHC-Ext without further tuning. Specifically, the IHC images in H&E2IHC-HER2 are processed using the 4B5 antibody, while the IHC images in H&E2IHC-Ext are stained with SP3 or CB11 antibodies. As shown in Table 20, performance generally drops compared to the in-domain setting (Table 2) due to the domain gap, as expected. Notwithstanding, the classifier trained on MagicStain-generated

Table 14: Evaluation of various methods on the MIST-Ki67 and MIST-PR datasets. *: $p < 0.05$ by the Wilcoxon signed-rank test for pairwise comparison with our method.

Method	MIST-Ki67				MIST-PR			
	FID↓	KID↓	PSNR↑	SSIM↑	FID↓	KID↓	PSNR↑	SSIM↑
Pix2pix	64.99	0.0441	15.21*±2.146	0.290*±0.101	75.11	0.0420	**15.41**±3.534	0.286*±0.105
Pix2pixHD	53.29	0.0331	14.59*±2.162	0.284*±0.100	60.08	0.0239	14.79*±3.563	0.276*±0.106
(Peng et al., 2024)	44.95	0.0223	14.64*±2.184	0.278*±0.100	65.80	0.0323	14.82*±3.520	0.262*±0.100
CycleGAN	38.77	0.0122	13.70*±1.908	0.252*±0.089	45.69	0.0113	13.40*±2.919	0.239*±0.086
AI-FFPE	36.98	0.0090	14.04*±1.946	0.275*±0.086	49.93	0.0193	13.48*±3.060	0.256*±0.089
USIGAN	27.93	0.0057	14.40*±1.876	0.273*±0.089	35.15	0.0064	14.22*±2.848	0.257*±0.090
ControlNet	250.3	0.2851	8.272*±1.324	0.170*±0.031	272.9	0.2609	7.585*±1.162	0.148*±0.028
HistDiST	73.67	0.0453	14.40*±2.130	0.275*±0.096	81.35	0.0452	14.33*±3.609	0.274*±0.103
StainFuser	63.15	0.0475	13.13*±1.376	**0.318**±0.107	72.06	0.0377	12.74*±2.514	**0.291**±0.130
pix2pix-Turbo/VIMs	161.5	0.1529	13.49*±1.458	0.232*±0.080	78.62	0.0436	14.89*±3.126	0.259*±0.095
PixCell	42.03	0.0179	14.56*±2.099	0.276*±0.098	45.28	0.0155	14.36*±3.353	0.269*±0.107
MagicStain (ours)	**27.81**	**0.0041**	**15.35**±2.305	0.305±0.109	**29.72**	**0.0015**	15.37±3.559	0.288±0.111

Table 15: Comparison between MagicStain models trained for 200 and 20 epochs on the H&E2IHC-HER2 dataset.

Epochs	FID↓	KID↓	PSNR↑	SSIM↑
200	31.76	0.0060	18.55±3.754	0.415±0.128
20	33.51	0.0033	18.58±3.668	0.411±0.125

images again outperforms those trained on images produced by all compared methods. In addition, its performance is slightly better than that of the classifier trained on real 4B5-antibody IHC images in H&E2IHC-HER2. This suggests that classifiers trained on high-quality virtually stained images may generalize better across antibody types, likely because virtual staining produces pathological representations that are less biased by antibody-specific staining variation. In conclusion, the promising downstream-task performance of MagicStain across both in- and out-of-domain evaluation indicates that the IHC images generated by MagicStain exhibit accurate HER2 expression, i.e., accurate pathological translation from H&E to IHC images.

B.14 Downstream Task Qualitative Analysis

In this section, we provide qualitative results for the downstream classification task. Figure 5 shows representative examples from Section 4 and Appendix B.13, with the corresponding classification results of various methods summarized in Table 21. Specifically, Cases 1–4 are drawn from the test set used in Section 4, while Cases 5–8 are from the H&E2IHC-Ext dataset evaluated in Appendix B.13. In HER2 scoring, "0" denotes no cancerous lesion, while "1+", "2+", and "3+" indicate increasing severity of cancerous lesions. As shown in Table 21, classifiers trained on images virtually stained by many existing SOTA methods struggle to accurately determine HER2 scores, especially for the "1+" and "2+" categories. In contrast, the classifier trained on MagicStain-generated images accurately recognizes all categories, demonstrating that the IHC images produced by MagicStain preserve accurate HER2 expression.

B.15 Additional Evaluation Metrics

To enable a more comprehensive pathological evaluation, we additionally employ a pathology-specific encoder—UNI2-h (Chen et al., 2024b)—to assess the gap between the generated results and the target domains by computing the mean cosine similarity of the feature embeddings extracted by UNI2-h for each pair on the H&E2IHC-HER2, BCI, and MIST datasets. We also replace the Inception network used for computing FID in Section 4 with UNI2-h (Chen et al., 2024b) such that the metric calculation is better aligned with the specialized field, following (Bhosale et al., 2025). As shown in Tables 22 and 23, MagicStain achieves the best UNI2-h similarity and UNI2-h-based FID scores on all these datasets, demonstrating its superior capability in clinically correct virtual staining and validating its effective design and strong performance under more pathology-relevant evaluation. These results are

Table 16: Training cost of various methods on the H&E2IHC-HER2 dataset.

Method	Training time (h)	Training VRAM (GB)	Trainable params (M)
PixCell (MLP)	5.10	12	24.166
PixCell (LoRA)	52.29	33	1.032
HistDiST	49.40	20	859.532
VIMs / pix2pix-Turbo	240.27	60	9.505
ControlNet	117.37	51	361.279
StainFuser	52.74	57	878.487
MagicStain (512 stage)	16.77	42	866.436
MagicStain (1024 stage)	8.13	55	8.633

Table 17: Inference cost for one 1024×1024 image.

Method	Inference time (s)	Inference VRAM (GB)
StainFuser	3.690	10.322
HistDiST	12.271	4.094
ControlNet	18.495	21.628
VIMs / pix2pix-Turbo	0.439	7.263
PixCell	4.457	9.060
MagicStain (ours)	**0.430**	**5.102**

consistent with the conclusions of our experiments on virtual staining and downstream tasks in the main text.

B.16 Failure Analysis

We further analyze a representative failure case of MagicStain (Figure 6). Here, a HER2 2+ region is mistakenly stained as 1+, with the moderate membrane staining intensity that defines this intermediate grade being underestimated. These intermediate grades are inherently difficult to learn: their membrane staining patterns are subtle and easily confused with adjacent grades, while paired training samples for them are comparatively scarce. Consequently, they remain harder to predict reliably than the clearly negative (0) and strongly positive (3+) grades, which we regard as a promising direction for future work.

C Additional WSI results

To further validate the generalizability of MagicStain on whole-slide images (WSIs), we additionally evaluate it on two H&E WSIs and two FFPE WSIs. These WSIs serve as an external test set, analogous to the H&E2IHC-Ext dataset in Appendix B.13. The methods compared are selected based on their performance in the quantitative and qualitative analyses in Section 4. We select the better-performing methods Pix2pixHD (Wang et al., 2018) and AI-FFPE (Ozyoruk et al., 2021) from the GAN-based methods, and pix2pix-Turbo (Parmar et al., 2024) from the diffusion-based ones. For the H&E WSIs, all compared methods (including ours) use models trained on the H&E2IHC-HER2 dataset for inference; for the FFPE WSIs, all methods use models trained on the FFPE2H&E dataset.

Because WSIs are ultra-high-resolution and direct inference on an entire WSI is infeasible, we partition each WSI into non-overlapping patches of 1024×1024 pixels. We perform inference on each patch and then stitch the patch outputs back together to reconstruct a full WSI. Due to the absence of paired target WSIs for metric computation (e.g., FID), we present the virtually stained WSIs to two board-certified pathologists for an overall, blind mean ranking on a scale of 1 (best) to 4 (worst), considering both image quality and pathology correctness.

The virtually stained WSIs are shown in Figures 7, 8, 9 and 10. The mean ranking of the four WSIs is: MagicStain (1.0), pix2pix-Turbo (2.0), Pix2pixHD (3.0), and AI-FFPE (4.0), demonstrating MagicStain's superior virtual staining quality when scaled up.

Table 18: Impact of enhancing the DAB channel on the H&E2IHC-HER2 dataset.

Variant	FID↓	KID↓	PSNR↑	SSIM↑
w DAB w H	33.18	0.0049	18.60±3.675	0.405±0.126
w DAB w/o H	37.78	0.0091	17.91±3.427	0.382±0.120
w/o DAB w H (Ours)	33.51	0.0033	18.58±3.668	0.411±0.125

Table 19: Impact of pretrained weights on the H&E2IHC-HER2 dataset.

Variant	FID↓	KID↓	PSNR↑	SSIM↑
w/o pretrain	37.12	0.0081	18.28±3.708	0.407±0.122
w pretrain (Ours)	33.51	0.0033	18.58±3.668	0.411±0.125

According to both pathologists' feedback, among all methods, MagicStain produces the most stable and realistic virtual staining results. More importantly, MagicStain achieves high visual quality with staining colors closely matching real HER2-IHC and H&E images, while preserving critical structure and pathology in the source input images. Although a mild mosaic ("checkerboard") artifact (Sun et al., 2023) is occasionally observed—mostly in background regions, it does not affect diagnostic interpretation. In contrast, all compared methods exhibit suboptimal staining with notable deficiencies, such as poor tissue structure continuity, blurred tumor cell boundaries, and incorrect cell pathology statuses—issues that substantially hinder accurate pathological assessment.

D Broader Impacts

AI-assisted pathology has attracted substantial research attention in recent years, driven by the potential to improve diagnostic consistency and reduce the workload of clinical pathologists. Virtual staining—digitally converting one histological stain to another—sits at the edge of this broader trend, offering a cost-effective route to multi-marker tissue analysis without repeated physical sectioning. Our work falls squarely within this domain and inherits its societal benefits: reducing reagent costs, enabling retrospective studies of archived slides, and potentially improving access to IHC-based diagnostics in resource-limited settings.

This work involves no clinical trials, no interventions with human subjects, and no new dataset releases beyond those already authorized by their respective sources. We foresee no potential for misuse, and the work raises no concerns related to privacy, security, discrimination, fairness, or research integrity. There are no conflicts of interest or external sponsorships associated with this research.

E Limitations and Future Work

In practical clinical settings, collecting paired cross-stain training data is typically time-consuming and costly, potentially restricting the scalability and broader adoption of MagicStain. Although paired data enhances virtual staining quality through direct correspondence, unpaired data offers much greater scalability due to its significantly larger availability. We plan to expand MagicStain by incorporating methods such as CycleGAN-Turbo (Parmar et al., 2024) in future work, which enable diffusion models to be trained on unpaired data.

Compared to VIMs, our MagicStain demonstrates effective virtual staining at a high resolution of 1024×1024 pixels, taking a step further toward the efficient re-staining of large-scale images, such as WSIs. Currently, due to the megabyte-sized nature of WSIs and hardware constraints, directly generating an entire WSI is infeasible for almost all virtual staining approaches. Existing methods often bypass this challenge by dividing a WSI into smaller patches, processing each patch individually, and stitching the outputs together (Asaf et al., 2024; Jewsbury et al., 2024; Koivukoski et al., 2023; Pati et al., 2024; Rana et al., 2018). However, this workaround often introduces noticeable "tiling" artifacts (Sun et al., 2023). In the future, we plan to realize tiling-artifact-free WSI virtual staining with MagicStain, by exploring diffusion path fusion techniques such as MultiDiffusion (Bar-Tal et al., 2023).

Table 20: Out-of-domain evaluation of a downstream classification task using the H&E2IHC-HER2 and H&E2IHC-Ext datasets.

Method	ACC↑	F1↑
H&E	0.4246	0.3205
Real IHC	0.7228	0.7298
Pix2pix	0.5930	0.5253
Pix2pixHD	0.6386	0.5430
(Peng et al., 2024)	0.7263	0.6795
CycleGAN	0.5263	0.4714
AI-FFPE	0.6912	0.6251
USIGAN	0.5825	0.5678
ControlNet	0.3965	0.3202
HistDiST	0.7228	0.6957
StainFuser	0.5930	0.5391
pix2pix-Turbo/VIMs	0.3509	0.2070
PixCell	0.6596	0.6249
w/o prior	0.7193	0.6884
w/o $\mathcal{L}_{\text{Expert}}$	0.7018	0.7188
MagicStain (ours)	**0.7404**	**0.7459**

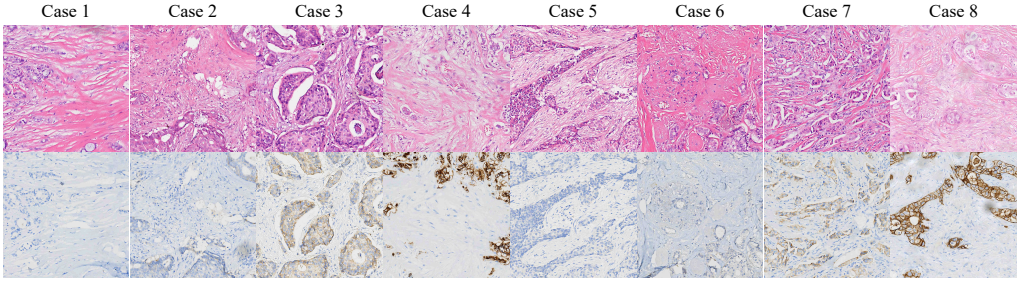

Figure 5: H&E and IHC image pairs from the H&E2IHC-HER2 (Cases 1–4) and H&E2IHC-Ext (Cases 5–8) datasets, with the IHC images to be classified in Table 21. HER2 scores: 0 indicates no cancerous lesion, while 1+, 2+, and 3+ represent increasing severity of cancerous lesions. Cases 1 to 4 (5 to 8) correspond to HER2 scores of 0, 1+, 2+, and 3+, respectively.

Table 21: Classification results for each case in Figure 5 by different classifiers. HER2 scores: 0 indicates no cancerous lesion, while 1+, 2+, and 3+ represent increasing severity of cancerous lesions. Cases 1 to 4 (5 to 8) correspond to HER2 scores of 0, 1+, 2+, and 3+, respectively. Bold texts indicate incorrect classifications.

Method	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7	Case 8
H&E	0	**3+**	**3+**	3+	**3+**	**3+**	**3+**	3+
Real IHC	0	1+	2+	3+	0	1+	2+	3+
Pix2pix	**1+**	1+	**0**	3+	**1+**	1+	**3+**	3+
Pix2pixHD	**1+**	1+	**0**	3+	0	1+	**3+**	3+
(Peng et al., 2024)	**1+**	1+	**3+**	3+	0	1+	**3+**	3+
CycleGAN	**2+**	1+	**3+**	**0**	**2+**	**0**	**0**	**2+**
AI-FFPE	**2+**	**0**	2+	3+	0	**0**	2+	3+
USIGAN	**1+**	1+	**0**	3+	**1+**	**2+**	2+	3+
ControlNet	0	**0**	**3+**	**0**	**3+**	**0**	**3+**	**0**
HistDiST	**2+**	**2+**	**0**	**2+**	0	**2+**	2+	**2+**
StainFuser	**3+**	**0**	2+	3+	0	**0**	**0**	3+
pix2pix-Turbo/VIMs	**1+**	**0**	**0**	**1+**	0	**0**	**0**	**1+**
PixCell	**1+**	1+	2+	3+	**1+**	**2+**	2+	3+
MagicStain (ours)	0	1+	2+	3+	0	1+	2+	3+

Table 22: UNI2-h-based FID and UNI2-h similarity of various methods on the H&E2IHC-HER2 and BCI datasets. *: $p < 0.05$ by the Wilcoxon signed-rank test for pairwise comparison with our method.

Method	H&E2IHC-HER2		BCI	
	FID↓	SIM↑	FID↓	SIM↑
Pix2pix	107.12	0.697*±0.0985	223.50	0.383*±0.1109
Pix2pixHD	109.53	0.672*±0.1050	212.35	0.397*±0.1253
(Peng et al., 2024)	80.06	0.735*±0.0891	161.60	0.464*±0.1290
CycleGAN	90.59	0.716*±0.0889	114.31	0.582*±0.1226
AI-FFPE	87.00	0.693*±0.1043	169.58	0.504*±0.1558
USIGAN	143.97	0.638*±0.0858	173.84	0.520*±0.1189
ControlNet	323.28	0.412*±0.1610	346.21	0.409*±0.1560
HistDiST	133.52	0.645*±0.1123	98.88	0.606*±0.1308
StainFuser	125.85	0.592*±0.1213	135.56	0.480*±0.1741
pix2pix-Turbo/VIMs	186.39	0.526*±0.0888	396.70	0.241*±0.0765
PixCell	111.95	0.661*±0.0958	80.95	0.646*±0.1063
MagicStain (ours)	**77.76**	**0.765**±0.0579	**66.52**	**0.688**±0.0999

Table 23: UNI2-h-based FID and UNI2-h similarity of various methods on the MIST datasets. *: $p < 0.05$ by the Wilcoxon signed-rank test for pairwise comparison with our method.

Method	MIST-HER2		MIST-ER		MIST-Ki67		MIST-PR	
	FID↓	SIM↑	FID↓	SIM↑	FID↓	SIM↑	FID↓	SIM↑
Pix2pix	164.89	0.411*±0.1249	160.64	0.425*±0.1249	170.98	0.467*±0.1210	166.58	0.397*±0.1138
Pix2pixHD	157.21	0.452*±0.1105	135.25	0.477*±0.1443	159.84	0.500*±0.1324	142.55	0.446*±0.1405
(Peng et al., 2024)	127.24	0.492*±0.1341	121.54	0.522*±0.1368	133.90	0.554*±0.1267	127.54	0.495*±0.1258
CycleGAN	132.01	0.564*±0.1369	91.69	0.620*±0.1278	104.17	0.633*±0.1118	99.76	0.601*±0.1334
AI-FFPE	133.10	0.534*±0.1268	138.35	0.536*±0.1235	111.53	0.575*±0.1233	100.41	0.569*±0.1272
USIGAN	75.46	0.637*±0.1259	71.59	0.660*±0.1300	67.82	0.674*±0.1185	74.75	0.650*±0.1387
ControlNet	278.37	0.417*±0.1397	252.71	0.403*±0.1483	231.41	0.671*±0.1199	266.91	0.364*±0.1387
HistDiST	127.39	0.512*±0.1465	128.80	0.526*±0.1501	130.36	0.472*±0.1683	125.50	0.520*±0.1444
StainFuser	90.45	0.512*±0.1420	88.02	0.516*±0.1742	106.78	0.549*±0.1416	94.18	0.502*±0.1580
pix2pix-Turbo/VIMs	232.42	0.410*±0.1129	170.97	0.446*±0.1434	212.89	0.451*±0.1060	151.52	0.501*±0.1268
PixCell	87.01	0.621*±0.1038	81.86	0.640*±0.1077	92.76	0.640*±0.1064	83.07	0.630*±0.1042
MagicStain (ours)	**69.84**	**0.670**±0.1144	**57.75**	**0.714**±0.1139	**63.44**	**0.715**±0.1038	**60.99**	**0.691**±0.1165

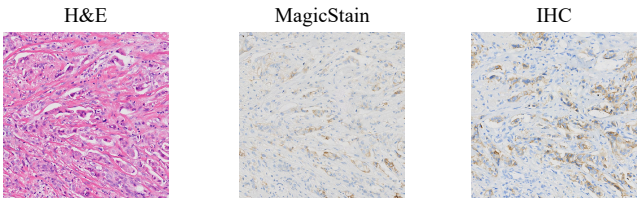

Figure 6: A representative failure case of MagicStain. A HER2 2+ region is mistakenly stained as 1+, with the moderate membrane staining intensity that characterizes this intermediate grade being underestimated.

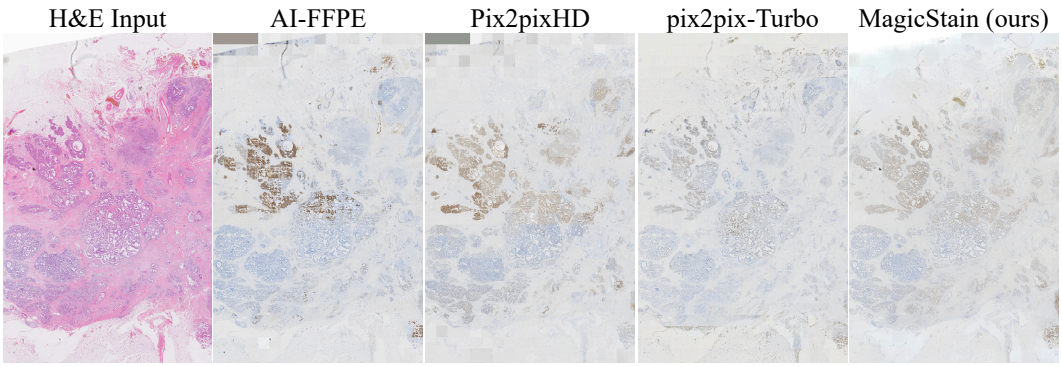

Figure 7: Virtual IHC WSIs stained from an H&E WSI input. Resolution: 19075×30000 pixels.

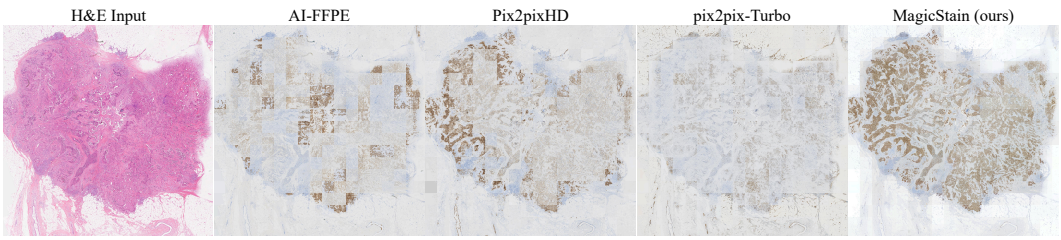

Figure 8: Virtual IHC WSIs stained from an H&E WSI input. Resolution: 18000×18000 pixels.

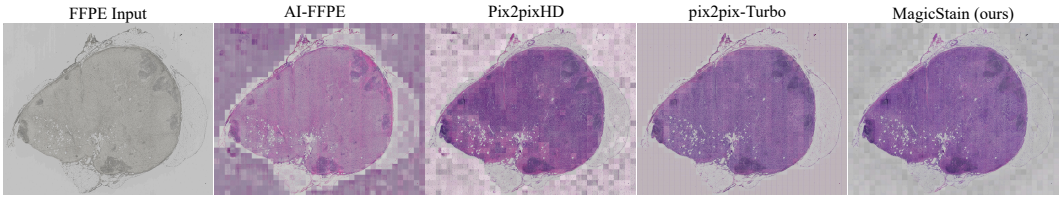

Figure 9: Virtual H&E WSIs stained from an FFPE WSI input. Resolution: 40000×32496 pixels.

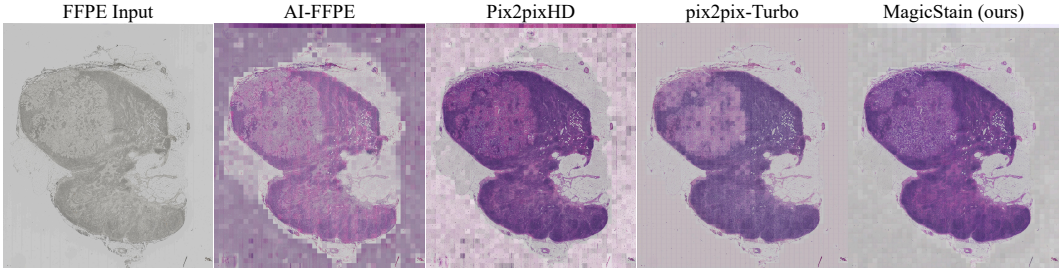

Figure 10: Virtual H&E WSIs stained from an FFPE WSI input. Resolution: 44000×51200 pixels.