ASP-DGRank: Attention-Guided Supervised Patch Ranking with Diffusion-Regularized GAN for Virtual Staining

Srijita Khatua[1] 24DR0190@IITISM.AC.IN
Manisha Verma[2] MANISHA@IITISM.AC.IN
Sudhakar Kumawat[3] SUDHAKAR@IITISM.AC.IN
[1,2,3] *IIT (ISM) Dhanbad*

Editors: Under Review for MIDL 2026

Abstract

Translating histopathological images across different staining modalities is a challenging task, as maintaining both structural consistency and biomarker integrity is critical for diagnostic reliability. To address this, we present ASP-DGRank, an attention-guided supervised patch-ranking model that integrates adversarial learning, Attention-guided Adaptive Supervised Patch (A^2SP) loss, and diffusion-based regularization. The proposed approach selects patches and adaptively weights them to mitigate the effect of noisy or misaligned regions, while emphasizing diagnostically relevant areas. We evaluate ASP-DGRank using the BCI dataset for the HER2 biomarker and the MIST dataset for four biomarkers (HER2, ER, PR, and Ki67). The method produces virtual stains that are structurally consistent and clinically meaningful. Quantitative analysis with structural and pathology-aware metrics indicates improvements compared to existing approaches. Overall, ASP-DGRank contributes to improving the reliability and interpretability of histopathological image translation.

Keywords: Adaptive Attention, Histopathological image translation, Diffusion-GAN, Biomarker, Virtual staining

1. Introduction

Cancer remains a leading cause of global mortality, creating significant challenges for healthcare systems and emphasizing the need for effective, affordable detection and treatment methods. Histopathological analysis, particularly Hematoxylin and Eosin (H&E) staining, is a key diagnostic tool for evaluating tissue abnormalities. While H&E staining is widely used and examined under a microscope, high-resolution whole slide imaging (WSI) has become more common, providing digital access to tissue morphology and disease features (Latonen et al., 2024). However, advanced cancer treatment often requires the evaluation of molecular biomarkers, such as Human Epidermal Growth Factor Receptor 2 (HER2), which is critical for breast cancer prognosis. Immunohistochemistry (IHC) is the standard method for HER2 detection, but is resource-intensive, limiting its accessibility in low-resource settings. In contrast, H&E staining is cost-effective and widely available, driving the development of computational methods like virtual staining, which can digitally replicate IHC staining, potentially accelerating diagnostics and improving access to advanced biomarker analysis (Klöckner et al., 2025). Recent advancements in deep learning, particularly through image-to-image translation with Generative Adversarial Networks (GANs), have enabled

the generation of virtual IHC stains. Existing methods are generally divided into supervised and unsupervised GAN-based approaches. Supervised methods, such as Pix2Pix (Liu et al., 2022) and Pyramid Pix2Pix (Isola et al., 2017), require tightly paired training data, which is challenging to obtain in histopathology due to issues like tissue deformation and staining variability. Unsupervised models, like CycleGAN (Zhu et al., 2017), avoid paired data but often struggle with pathological fidelity due to the non-invertible nature of histological staining. Models like CUT (Park et al., 2020) and ASP (Li et al., 2023) improve upon this by integrating contrastive learning, though they still rely on weakly paired data. Diffusion models, which have shown strong results in other imaging domains (Klöckner et al., 2025; Lin et al., 2025), are emerging as a potential alternative for virtual staining, but their application to IHC is still limited, with only the PST-DIFF model (He et al., 2024) available for the BCI dataset.

Despite these advances, two key challenges remain. Firstly, current models struggle to jointly preserve molecular biomarker expression and fine tissue morphology essential for diagnostic reliability. Secondly, many methods perform well on a single dataset or biomarker (e.g., HER2) but generalize poorly to others, such as Ki67, ER, and PR, with few showing robust performance across datasets and markers (Klöckner et al., 2025; Lin et al., 2025; Latonen et al., 2024). Additionally, supervised methods require strictly paired data and accurate registration, both computationally demanding and prone to error; while unsupervised approaches avoid registration but typically require longer training and offer less reliable feature preservation.

To address these limitations, we introduce **ASP-DGRank**, an attention-guided adaptive supervised patch-ranking model that integrates adversarial learning with diffusion-based regularization. The framework is thoughtfully designed to preserve structural fidelity and pathological relevance while remaining adaptable across various datasets and biomarkers. Notably, it supports scenarios wherein pairs of consecutive tissue slices are not perfectly aligned, yet are expected to share corresponding diagnostic information at the patch level. The main contributions of this work include:

- An attention module that selects patches and adaptively weights, followed by Attention-guided Adaptive Supervised Patch (A^2SP) loss, emphasizing diagnostically important regions and reducing the impact of noisy or misaligned areas, leveraging paired or weakly paired data.

- A diffusion module that corrupts real and generated samples with timestep-indexed noise, enhancing structural consistency in virtual IHC stains and stabilizing GAN training across multiple biomarkers.

- A robust framework that jointly maintains molecular precision and tissue morphology, leading to more reliable and interpretable virtual staining while ensuring inconsistent patches contribute less during learning.

- Experimental validation showing improved structural integrity across datasets and biomarkers, with consistent gains across diverse quantitative evaluation metrics, including both paired and unpaired metrics.

2. Related Work

Virtual staining, which involves translating images between different staining modalities, typically relies on supervised and unsupervised learning methods. The most widely adopted are Generative Adversarial Networks (GANs) (Zhu et al., 2017), often combined with contrastive learning techniques such as Contrastive Unpaired Translation (CUT) (Park et al., 2020) to improve translation quality when paired data are unavailable. More recently, Diffusion Models have attracted growing attention, showing strong performance in virtual staining tasks and already surpassing GANs in other image translation domains (Klöckner et al., 2025; Lin et al., 2025).

2.1. Structural Preservation

Early supervised approaches, such as Pix2Pix (Isola et al., 2017) and Pyramid Pix2Pix (Liu et al., 2022), require pixel-level aligned H&E and IHC images. Pix2Pix combines adversarial and reconstruction losses with the source image as conditional input, while Pyramid Pix2Pix extends this by employing multi-scale processing and L1 optimization at different resolutions. Despite their effectiveness, these methods are constrained by the practical difficulty of collecting perfectly aligned training pairs.

Unsupervised models overcome this limitation by learning mappings between staining domains without explicit correspondence. CycleGAN (Zhu et al., 2017) enforces cycle consistency to maintain correspondence between input and reconstructed images, while CUT (Park et al., 2020) introduces contrastive learning to strengthen this mapping further. Adaptive Supervised PatchNCE (ASP) (Li et al., 2023) builds on CUT by incorporating adaptive contrastive loss, making training more robust against poorly matched image pairs. Although these methods effectively preserve tissue structure and anatomical integrity, they often overlook subtle pathological cues, such as molecular markers or disease-specific tissue properties, which limit their diagnostic utility.

2.2. Pathological Preservation

To address this gap, recent methods have focused on preserving pathological semantics. PSPStain (Chen et al., 2024) enhances semantic alignment between H&E and IHC images and demonstrates strong performance at the molecular level, though its generalizability remains uncertain due to validation on a limited biomarker (e.g., HER2-stain). ConfusionGAN (Li et al., 2024) further improves diagnostic accuracy by integrating a multi-branch discriminator with a patch-level pathology information extractor (PPIE). Applied to hepatocellular carcinoma, it produces highly realistic IHC translations and boosts instance-level classification when paired with multiple instance learning (MIL) techniques. This framework also introduced a high-resolution dataset of paired H&E and GPC3-stained images, enabling broader evaluation.

3. Methodology

In this framework, given an input pair consisting of an H&E image $I \in \mathbb{R}^{H\times W\times C}$ and its corresponding label i.e., real IHC $K^R \in \mathbb{R}^{H\times W\times C}$, the generator backbone produces a synthesized IHC image, i.e., fake IHC $K^F = G(I) \in \mathbb{R}^{H\times W\times C}$. The generator G has

an encoder-decoder based architecture with L intermediate layers. We denote the encoder feature map by $f_\ell(X) \in \mathbb{R}^{C_\ell \times H_\ell \times W_\ell}$, at layer $\ell \in \{0, \ldots, L-1\}$ for any image $X \in \mathbb{R}^{H \times W \times C}$, where C_ℓ is the number of channels and H_ℓ, W_ℓ are the spatial dimensions at layer ℓ. For a predefined set of contrastive layers $\mathcal{L} = \{0, 4, 8, 12, 16\}$, we extract encoder features $\mathcal{F}_I = \{f_\ell(I)\}_{\ell \in \mathcal{L}}$, $\mathcal{F}_K^F = \{f_\ell(K^F)\}_{\ell \in \mathcal{L}}$, $\mathcal{F}_K^R = \{f_\ell(K^R)\}_{\ell \in \mathcal{L}}$ for the input H&E image, the generated IHC, and the real IHC, respectively. A separate projection network F operates on each feature map to obtain patch-level embeddings. For a given layer ℓ, F_ℓ samples N spatial locations from $f_\ell(X)$ and projects the corresponding feature vectors into a d-dimensional embedding space: $F_\ell : \mathbb{R}^{C_\ell \times H_\ell \times W_\ell} \to \mathbb{R}^{K \times d}$; resulting the patch sets $Z_I^\ell = F_\ell(f_\ell(I))$, $Z_K^{F,\ell} = F_\ell(f_\ell(K^F))$, $Z_K^{R,\ell} = F_\ell(f_\ell(K^R))$ respectively and collect them across $\ell \in \mathcal{L}$. By using a shared encoder of G for I, K^F, and K^R, the feature sets $\mathcal{F}_I$, $\mathcal{F}_K^F$, and $\mathcal{F}_K^R$ all inhabit a common latent space. This shared representation is essential for defining patch-level correspondences across domains and for expressing both global content preservation and attention-guided supervision as contrastive objectives on the same set of encoder features, as shown in Figure 1.

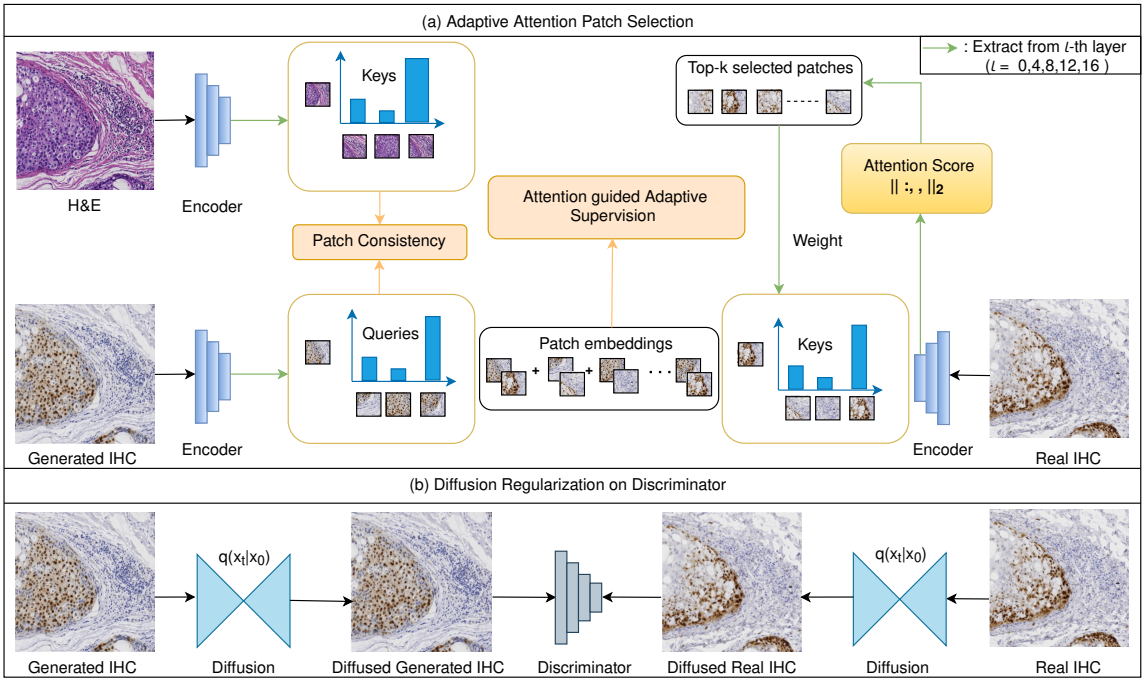

Figure 1: The architecture of the proposed ASP-DGrank. Important modules : (a) Attention module tasked to select patches and perform adaptive supervision, (b) Diffusion module conditioned on the discriminator stabilizes the GAN training.

3.1. Attention-guided patch selection

We derive an attention map at each layer ℓ directly from the feature magnitudes from the encoder features of the real IHC image $\mathcal{F}_K^R = \{f_\ell(K^R)\}_{\ell \in \mathcal{L}}$. For a feature map $f_\ell(K^R) \in \mathbb{R}^{C_\ell \times H_\ell \times W_\ell}$, the attention score at spatial location (h, w) is defined as the ℓ_2-norm across

channels,

$$A_\ell(h,w) = \left\| f_\ell(K^R)_{:,h,w} \right\|_2 = \left(\sum_{c=1}^{C_\ell} \left(f_\ell(K^R)_{c,h,w} \right)^2 \right)^{1/2}, \tag{1}$$

which yields an attention map $A_\ell \in \mathbb{R}^{H_\ell \times W_\ell}$. Then we reshape A_ℓ to a vector of length $H_\ell W_\ell$ and use it as a discrete sampling distribution over normalized scores to select informative patches. Using A_ℓ, we sample a set of spatial indices $\mathcal{S}_\ell$ of size N via multinomial sampling over the $H_\ell W_\ell$ locations. The projection network F_ℓ then extracts patch embeddings at these locations for both the real and fake IHC feature maps: $\{u_{\ell,s}\}_{s\in\mathcal{S}_\ell} = F_\ell\big(f_\ell(K^R)\big)$, $\{v_{\ell,s}\}_{s\in\mathcal{S}_\ell} = F_\ell\big(f_\ell(K^F)\big)$ respectively. Importantly, the same index set $\mathcal{S}_\ell$ is reused for real and fake features, ensuring that $u_{\ell,s}$ and $v_{\ell,s}$ are extracted from corresponding spatial locations.

In our framework, we define A^{2}SP as the Adaptive Supervised PatchNCE loss applied after attention-guided sampling. The attention maps A_ℓ, computed solely from the real IHC encoder features $\{f_\ell(K^R)\}_{\ell\in\mathcal{L}}$, determine the informative spatial locations $\mathcal{S}_\ell$ at each layer ℓ. At these locations, we obtain paired embeddings $\{u_{\ell,s}, v_{\ell,s}\}_{s\in\mathcal{S}_\ell}$ for real and fake IHC. Them we define the InfoNCE loss $\ell_{\text{InfoNCE}}(u_{\ell,s}, v_{\ell,s}; \tau)$ and assign to each pair an adaptive weight $\theta_{\ell,s}$ that depends on both the training progress and the current similarity, following (Li et al., 2023). After normalization over $\mathcal{S}_\ell$, the attention-guided A^{2}SP loss at layer ℓ is

$$\mathcal{L}^\ell_{\text{A}^2\text{SP}} = \frac{1}{N} \sum_{s\in\mathcal{S}_\ell} \theta_{\ell,s}\, \ell_{\text{InfoNCE}}(u_{\ell,s}, v_{\ell,s}; \tau) \tag{2}$$

and the overall A^{2}SP loss is obtained by averaging over all selected layers:

$$\mathcal{L}_{\text{A}^2\text{SP}} = \frac{1}{|\mathcal{L}|} \sum_{\ell\in\mathcal{L}} \mathcal{L}^\ell_{\text{A}^2\text{SP}} \tag{3}$$

where N is the number of sampled patches per layer . Here A^{2}SP can be interpreted as an attention-guided ASP loss (Li et al., 2023) as the attention maps first select discriminative patches in the real IHC, and ASP then imposes an adaptively weighted contrastive alignment between the corresponding fake and real patch embeddings at those locations.

3.2. Diffusion Regularization

To regularize the adversarial learning process, we corrupt real and generated IHC patches using a forward diffusion process before passing them to the discriminator. We consider a discrete-time diffusion process defined by a schedule of noise variances $\{\beta_t\}_{t=1}^T$, with $\alpha_t = 1-\beta_t$ and cumulative products $\bar{\alpha}_t = \prod_{p=1}^t \alpha_p$. The corresponding forward distribution at time t for input image $x_0 \in \mathbb{R}^{H\times W\times C}$ is

$$q(x_t \mid x_0) = \mathcal{N}\big(x_t\,;\, \sqrt{\bar{\alpha}_t}\, x_0,\ (1-\bar{\alpha}_t)\,\mathbf{I}\big), \qquad t \in \{1,\dots,T\}. \tag{4}$$

Sampling from this distribution can be written as

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0,\mathbf{I}), \tag{5}$$

where the sequence $\{\beta_t\}$ (and hence $\{\bar{\alpha}_t\}$) is chosen according to a predefined schedule. For each training step, we first draw a diffusion time t from a data-dependent distribution over $\{1, \ldots, T\}$, biased towards later steps to emphasize stronger corruptions. We then sample a diffused image x_t according to $q(x_t \mid x_0)$. When applied to the real and generated IHC images, this yields pairs $(\tilde{K}^F, t^F) \sim q(\cdot \mid K^F)$, $(\tilde{K}^R, t^R) \sim q(\cdot \mid K^R)$, where $\tilde{K}^F = x_{t^F}$ and $\tilde{K}^R = x_{t^R}$ are diffused versions of the fake and real images at randomly sampled diffusion times t^F, t^R. A separate discriminator network $\mathcal{D}$ is then conditioned on both the diffused image and its diffusion time, and produces a spatial realism score. We adopt a least-squares adversarial objective on these diffusion-conditioned inputs. The discriminator loss is

$$\mathcal{L}_{\text{disc}} = \frac{1}{2}\Big(\mathbb{E}\big[(\mathcal{D}(\tilde{K}^F, t^F) - 0)^2\big] + \mathbb{E}\big[(\mathcal{D}(\tilde{K}^R, t^R) - 1)^2\big]\Big), \tag{6}$$

while the generator adversarial loss is

$$\mathcal{L}_{\text{GAN}} = \mathbb{E}[(\mathcal{D}(\tilde{K}^F, t^F) - 1)^2]. \tag{7}$$

By evaluating real and generated samples along a stochastic diffusion trajectory, rather than only in the clean image space, the discriminator learns to separate the two distributions across a continuum of noise levels. This diffusion-based regularization smooths the discriminator's decision boundary, providing more stable and informative gradients to guide the generator.

In addition, we employ the PatchNCE loss, as introduced in CUT (Park et al., 2020), to encourage the generator to preserve local content by maximizing the mutual information between corresponding patches of the input H&E image and the generated IHC image. To further enforce consistency of low-frequency structure, we incorporate the Gaussian pyramid reconstruction loss (Liu et al., 2022) that penalizes discrepancies between real and generated IHC images across multiple spatial scales. Finally, the overall training objective of our generator is defined as follows:

$$\mathcal{L}_G = \lambda_{\text{GAN}}\,\mathcal{L}_{\text{GAN}} + \lambda_{\text{P}}\,\mathcal{L}_{\text{PatchNCE}} + \lambda_{\text{A}^2\text{SP}}\,\mathcal{L}_{\text{A}^2\text{SP}} + \lambda_{\text{GP}}\,\mathcal{L}_{\text{GP}}, \tag{8}$$

where $\mathcal{L}_{\text{PatchNCE}}$ is the patch-wise contrastive content loss and $\mathcal{L}_{\text{GP}}$ is the Gaussian pyramid reconstruction loss.

4. Experiments

4.1. Datasets

In our experiments, we utilized the following datasets, namely, BCI (Liu et al., 2022) and MIST (Li et al., 2023). The BCI dataset includes various levels of HER2 expression, while the MIST dataset contains IHC staining data for HER2, PR, ER, and Ki67, with both datasets divided into training and testing sets, as shown in Table 1.

4.2. Training Details

The model is trained on an NVIDIA RTX 5000 GPU. We adopt the CUT training protocol with a batch size of 1, where input images are randomly cropped to a spatial resolution of

Table 1: The number of image pairs in each dataset.

Dataset	BCI_{HER2}	$MIST_{HER2}$	$MIST_{ER}$	$MIST_{PR}$	$MIST_{Ki67}$
Training Set	3896	4642	4153	4139	4361
Testing Set	977	1000	1000	1000	1000

512×512. The network is optimized for 40 epochs with a fixed initial learning rate of 0.0002, followed by 5 epochs of linear decay. Optimization is performed using the Adam optimizer with primary decay rates of 0.5 and 0.999, respectively. The number of sampled patches per layer is set to $N = 256$. The weighting coefficients for the different loss components are chosen as $\lambda_{GAN} = 1.0$, $\lambda_{P} = 10.0$, $\lambda_{GP} = 10.0$, $\lambda_{A^2SP} = 10.0$ with the PatchNCE temperature parameter fixed to $\tau = 0.07$. For the diffusion module, we use noise schedules $\beta_{start} = 10^{-4}$ and $\beta_{end} = 2 \times 10^{-2}$, and $\sigma = 0.05$ followed by diffusion timesteps T adapted between $t_{min} = 10$ and $t_{max} = 1000$.

4.3. Evaluation Metrics

We evaluate the similarity between real and generated images using various metrics. The Structural Similarity Index Measure (SSIM) (Wang et al., 2004) quantifies the structural similarity between generated and real images by analyzing luminance, contrast, and texture information. The Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018) metric is utilized to provide a perceptual quality measure, leveraging deep learning features to assess image similarity in alignment with human visual perception. For perceptual and distributional similarity, the Fréchet Inception Distance (FID) and Kernel Inception Distance (KID) (Mentzer et al., 2020) assess the divergence between the feature distributions of real and generated images in the Inception embedding space. The Peak Signal-to-Noise Ratio (PSNR) (Tanchenko, 2014) measures reconstruction fidelity based on pixel-level intensity differences, where higher values indicate better image quality. The average pathological heatmap variance (PHV_{avg}) (Liu et al., 2021) quantifies the preservation of critical pathological features, ensuring the fidelity of generated images in terms of their medical relevance. The average correlation (R_{avg}) (Liu et al., 2021), derived from content-preserving correlation for structural fidelity with H&E images and pathology-preserving correlation for diagnostic information consistency with IHC images, captures both content and pathological alignment during the image translation process.

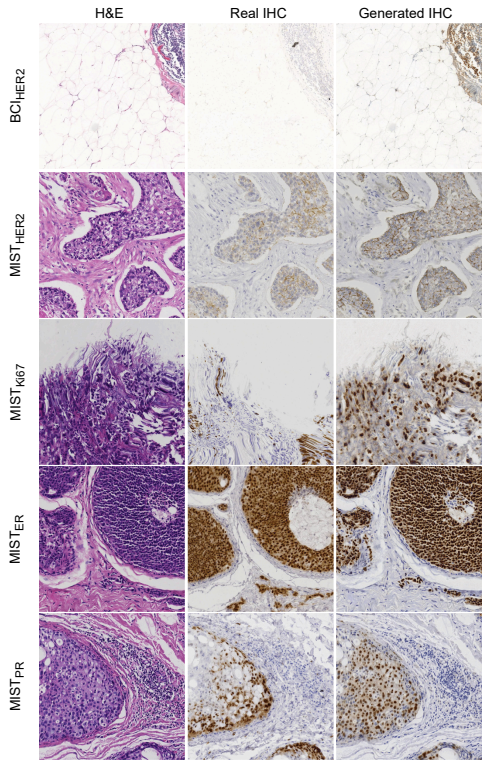

Figure 2: Visualization of our method on BCI and MIST dataset.

5. Results and Comparison

To assess the effectiveness of ASP-DGRank, we first performed experiments on the BCI and MIST datasets for the HER2 biomarker. Table 2 and Table 3 report that our model consistently achieved better results across all metrics, with the highest SSIM values highlighting its ability to preserve structural details and intensity patterns while ensuring reliable stain translation across the HER2 biomarker for the paired setting.

Table 2: Quantitative comparison of virtual staining on BCI_{HER2}. The values of the existing models are reported from (He et al., 2024). The **best** and second best values are highlighted.

Methods	PSNR ↑	SSIM ↑
SynDiff (Özbey et al., 2023)	14.28±2.52	0.32±0.14
Pix2pix GAN (Isola et al., 2017)	12.62±3.24	0.21±0.09
CUT (Park et al., 2020)	15.32±4.12	0.35±0.14
StainGAN (Shaban et al., 2019)	16.24±5.21	0.39±0.15
RegGAN (Kong et al., 2021)	18.22±5.51	0.37±0.18
Unit (Liu et al., 2017)	15.75±4.19	0.33±0.18
PST-Diff (He et al., 2024)	16.75±4.20	0.38±0.11
ASP-DGrank (paired)	**20.43±14.41**	**0.60±0.01**
ASP-DGrank (unpaired)	17.40±32.50	0.61±0.01

In addition to the primary evaluation, we also assessed our model for an unpaired setting, following CUT (Park et al., 2020), to demonstrate its efficacy in handling weakly paired data. For the unpaired setting, we use the same datasets but deliberately discard the explicit image-level pairing, drawing H&E and IHC images independently from their respective domains so that only weak, dataset-level correspondence is retained, whereas in the paired configuration each H&E image is presented together with its corresponding IHC image from the same or closely aligned tissue section, providing approximate spatial alignment for supervised training. This evaluation was conducted to further validate the versatility of our approach, showing that it can perform effectively even when explicit paired data is not available.

To comprehensively assess the robustness of our proposed method, we extended the evaluation to additional IHC biomarkers within the MIST dataset. As summarized in Table 4, ASP-DGRank consistently demonstrated strong performance compared to existing state-of-the-art methods. Figure 2 provides the qualitative analysis of our proposed method.

6. Ablation Studies

To analyze the contribution of each component in ASP-DGRank, we performed an ablation study on the $MIST_{HER2}$ dataset. The baseline model is ResNet-6Blocks as the generator and a 5-layer PatchGAN as the discriminator without attention or diffusion modules. We first added the diffusion module (DM) to the baseline, conditioned on the discriminator. The DM improves structural consistency and stabilizes GAN training through a diffusion-based reconstruction process. Next, we introduced only the attention module (AM), which selec-

Table 3: Quantitative comparison on the MIST_{HER2} dataset. The values of the existing models are reported from (Wang et al., 2025). KID values multiplied by 1000 are shown. The **best** and second best values are highlighted.

Methods	KID ↓	FID ↓	LPIPS ↓	SSIM ↑
Pix2pix (Isola et al., 2017)	102.0	147.6	0.470	0.1981
PyramidP2P (Liu et al., 2022)	80.3	180.5	0.467	0.1979
TDKStain (Peng et al., 2024)	43.0	138.3	0.460	0.1617
PSPStain (Chen et al., 2024)	34.0	148.1	0.446	0.1738
ASP (Li et al., 2023)	23.9	89.3	0.455	0.2004
ODA-GAN (Wang et al., 2025)	**8.6**	68.0	0.420	0.1893
ASP-DGrank (paired)	17.2	**50.1**	**0.399**	**0.3451**
ASP-DGrank (unpaired)	13.9	56.1	0.411	0.2883

Table 4: Quantitative comparison of virtual staining on MIST_{Ki67}, MIST_{PR}, and MIST_{ER}. The values of the existing models are reported from (Li et al., 2023). KID values multiplied by 1000 are shown. The **best** and second best values are highlighted.

Methods	MIST_{Ki67}			MIST_{PR}			MIST_{ER}		
	SSIM ↑	FID ↓	KID ↓	SSIM ↑	FID ↓	KID ↓	SSIM ↑	FID ↓	KID ↓
CycleGAN (Zhu et al., 2017)	**0.3875**	343.9	317.9	0.2232	96.1	96.6	0.1982	125.7	95.1
CUT+LGP (Park et al., 2020)	0.1909	76.1	43.5	0.2153	54.6	20.1	0.2217	43.7	8.7
Pix2Pix (Isola et al., 2017)	0.1819	147.0	142.4	0.1617	183.8	148.1	0.1500	128.1	79.0
PyramidP2P (Liu et al., 2022)	0.2286	94.4	78.0	0.2403	98.8	59.5	0.2172	107.4	84.2
ASP (Li et al., 2023)	0.2410	51.0	19.1	0.2159	**44.8**	10.2	0.2061	**41.4**	**5.8**
ASP-DGrank (paired)	0.3554	**45.22**	**10.5**	**0.3629**	45.03	**8.8**	**0.3459**	58.04	22.8

tively emphasizes tissue patches containing critical pathological information. By adaptively focusing on diagnostically important regions, the AM enhances local feature representation while maintaining global structure. Finally, we combined both modules within the baseline model. This integration provided the best trade-off between visual fidelity and structural similarity, achieving the lowest FID score alongside competitive SSIM values. The complete results of this ablation study are summarized in Table 5, highlighting the complementary role of the attention and diffusion modules in improving IHC stain translation.

Table 5: Ablation studies on $\text{MIST}_{\text{HER2}}$. Here, DM and AM refer to the Diffusion Module and Attention module, respectively. KID values multiplied by 1000 are shown.

Setting			$\text{MIST}_{\text{HER2}}$			
Baseline	DM	AM	SSIM ↑	FID ↓	KID ↓	PSNR ↑
✓	×	×	0.3097	85.85	40.0	13.8966
✓	✓	×	0.3298	62.13	23.6	14.6597
✓	×	✓	0.3384	60.71	21.8	14.8234
✓	✓	✓	0.3451	50.12	17.2	14.9124

According to Figure 3, our method produces virtual IHC images with a more pronounced and spatially coherent DAB signal, alongside improved preservation of underlying tissue architecture, compared to the baseline variants. In particular, cell-rich regions and biomarker-

positive structures are rendered with sharper boundaries and more realistic contrast, while background regions remain comparatively clean and less corrupted by spurious staining.

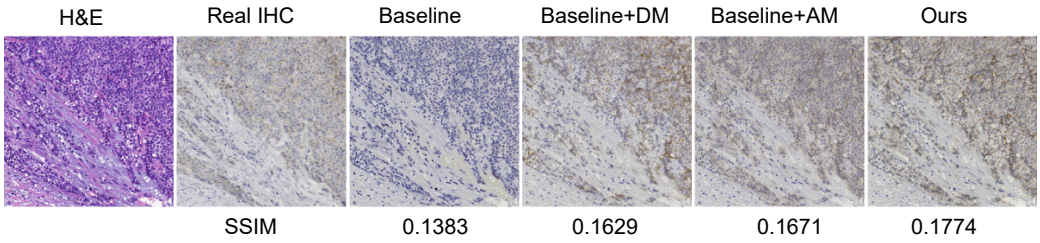

Figure 3: Ablation studies on $\text{MIST}_{\text{HER2}}$. Here, DM and AM refer to the Diffusion Module and Attention module, respectively.

7. Conclusion

Comprehensive experiments on the BCI and MIST datasets show that ASP-DGRank consistently outperforms existing methods in terms of structural similarity, perceptual quality, and feature preservation, with particularly strong performance on the BCI_{HER2} cohort. Unlike many state-of-the-art approaches that struggle to generalize across datasets and biomarkers, our model demonstrates robust and reliable performance in both paired and unpaired settings. By prioritizing perceptual fidelity and pathology-aware feature alignment, ASP-DGRank produces results that are more relevant for diagnostic applications.

Limitations: This study has several limitations. First, our model performs best on paired datasets and also works effectively on weakly paired datasets, where consecutive tissue slices may not be perfectly aligned, yet still share corresponding diagnostic information at the patch level. However, the method may not be suitable for fully unpaired data, where this correspondence is absent, and the underlying assumptions of the framework are not fulfilled. Second, the evaluation is limited to the paired datasets, namely BCI and MIST; therefore, the model's generalizability to other staining methods or cancer types remains to be validated. Finally, automated IHC synthesis is limited by biological variability, domain shifts, staining inconsistencies, and metrics that may miss subtle pathological details.

References

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*, pages 384–394. Springer, 2024.

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*, 2024.

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*, pages 1125–1134, 2017.

Pascal Klöckner, José Teixeira, Diana Montezuma, João Fraga, Hugo M Horlings, Jaime S Cardoso, and Sara P Oliveira. H&E to IHC virtual staining methods in breast cancer: an overview and benchmarking. *npj Digital Medicine*, 8(1):384, 2025.

Lingke Kong, Chenyu Lian, Detian Huang, Yanle Hu, Qichao Zhou, et al. Breaking the dilemma of medical image-to-image translation. *Advances in Neural Information Processing Systems*, 34:1964–1978, 2021.

Leena Latonen, Sonja Koivukoski, Umair Khan, and Pekka Ruusuvuori. Virtual staining for histology by deep learning. *Trends in Biotechnology*, 42(9):1177–1191, 2024.

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*, pages 632–641. Springer, 2023.

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*, pages 11259–11268, 2024.

Weiping Lin, Yihuang Hu, Runchen Zhu, Baoshun Wang, and Liansheng Wang. Virtual staining for pathology: Challenges, limitations and perspectives. *Intelligent Oncology*, 2025.

Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. *Advances in neural information processing systems*, 30, 2017.

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*, pages 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.

Fabian Mentzer, George D Toderici, Michael Tschannen, and Eirikur Agustsson. High-fidelity generative image compression. *Advances in neural information processing systems*, 33:11913–11924, 2020.

Muzaffer Özbey, Onat Dalmaz, Salman UH Dar, Hasan A Bedel, Şaban Özturk, Alper Güngör, and Tolga Cukur. Unsupervised medical image translation with adversarial diffusion models. *IEEE Transactions on Medical Imaging*, 42(12):3524–3539, 2023.

Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European conference on computer vision*, pages 319–345. Springer, 2020.

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*, pages 3–13. Springer, 2024.

M Tarek Shaban, Christoph Baur, Nassir Navab, and Shadi Albarqouni. Staingan: Stain style transfer for digital histological images. In *2019 Ieee 16th international symposium on biomedical imaging (Isbi 2019)*, pages 953–956. IEEE, 2019.

Alexander Tanchenko. Visual-psnr measure of image quality. *Journal of Visual Communication and Image Representation*, 25(5):874–878, 2014.

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*, pages 25920–25929, 2025.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

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*, pages 586–595, 2018.

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*, pages 2223–2232, 2017.

Appendix A. Additional quantitative evaluations of ASP-DGrank model for paired setting

Table 6: Quantitative Metrics of virtual staining on BCI and MIST dataset for paired setting. For metrics with ↑, higher values are better, while for metrics with ↓, lower values are better. KID values multiplied by 1000 are shown.

Dataset	SSIM ↑	PSNR ↑	FID ↓	KID ↓	LPIPS ↓	PHV_{avg} ↓	R_{avg} ↑
BCI_{HER2}	0.6012	20.4317	55.02	13.5	0.3380	0.4408	0.7278
$MIST_{HER2}$	0.3451	14.9124	50.12	17.2	0.3998	0.5167	0.7143
$MIST_{Ki67}$	0.3554	14.7880	45.22	10.5	0.3820	0.5070	0.7909
$MIST_{PR}$	0.3629	14.4624	45.03	8.8	0.3709	0.5039	0.8304
$MIST_{ER}$	0.3459	14.052	58.04	22.8	0.3729	0.4983	0.7961

Appendix B. Additional quantitative evaluations of ASP-DGrank model for unpaired setting

Table 7: Quantitative Metrics of virtual staining on BCI_{HER2} and $MIST_{HER2}$ for unpaired setting. For metrics with ↑, higher values are better, while for metrics with ↓, lower values are better. KID values multiplied by 1000 are shown.

Dataset	SSIM ↑	PSNR ↑	FID ↓	KID ↓	LPIPS ↓	PHV_{avg} ↓	R_{avg} ↑
BCI_{HER2}	0.6177	17.4034	52.92	11.1	0.4276	0.4822	0.6601
$MIST_{HER2}$	0.2883	13.8958	56.19	13.9	0.4110	0.5274	0.5750