PIXSTAIN: PIXEL-SPACE DIFFUSION WITH CROSS-VIEW REPRESENTATION ALIGNMENT FOR VIRTUAL STAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

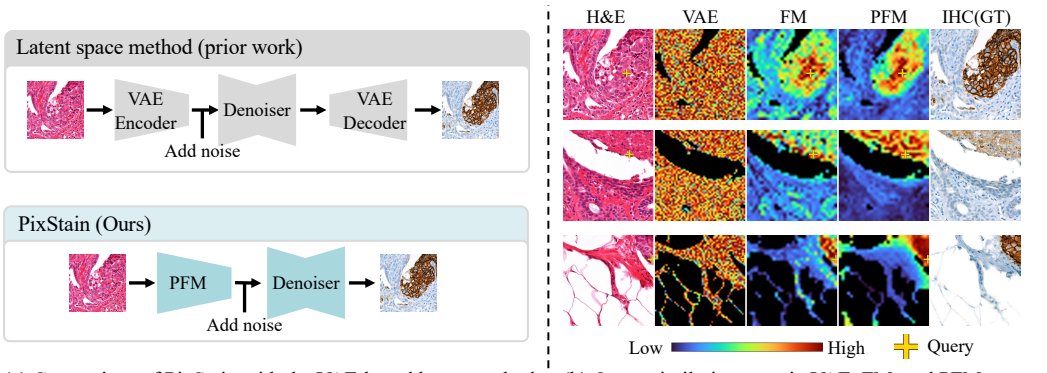
Virtual staining translates digital histology images between staining modalities, offering a faster and less expensive alternative to immunohistochemistry (IHC) and multiplex immunofluorescence (mIF). Existing diffusion-based virtual staining approaches generate high-quality images, but they operate in a VAE latent space that discards fine histological detail, require iterative sampling, and rely on output-space supervision alone. We observe that pathology foundation model (PFM) features provide informative spatial priors for virtual staining: their similarity maps closely align with regions of high marker expression, whereas those derived from VAE latents are dominated by noise. We therefore propose PixStain, a pixel-space diffusion model for virtual staining built on cross-view representation alignment. PixStain conditions a pixel-space Vision Transformer on multi-level PFM features of the H&E image and trains it with flow matching under x-prediction, so that a single stage of training yields the target marker image in a single inference step. Cross-view alignment matches shallow student features from spatially disjoint marker noising and H&E masking to deeper teacher features from a cleaner view. These complementary corruptions encourage the student to integrate information from both inputs. On three publicly available same-section datasets covering H&E-to-IHC, H&E-to-mIHC, and H&E-to-mIF, PixStain achieves state-of-the-art performance across pixel-level and perceptual metrics, as well as cell classification and segmentation metrics, except for pixel-level metrics on HEMIT, where it ranks second.

1 INTRODUCTION

Histological staining underpins disease diagnosis and life-science research (Bancroft & Gamble, 2008). Hematoxylin and eosin (H&E), the most widely used stain and the gold standard of pathological diagnosis, renders nuclei and extracellular matrix in contrasting colors, but it reveals tissue morphology rather than protein marker expression (Ali et al., 2017). Immunohistochemistry (IHC) uses antigen–antibody binding to reveal the expression of individual protein markers, making it one of the most widely used assays in clinical pathology. A single marker, however, cannot characterize the complex tumor microenvironment (TME) (Lin et al., 2023). Multiplex immunofluorescence (mIF) visualizes and quantifies several protein markers at once, modeling the TME more faithfully. Both IHC and mIF, however, remain costly and slow to acquire.

Virtual staining, which translates a digital slide from one staining modality into another, has therefore attracted broad interest as a cheaper and faster alternative. Existing methods fall into three families: regression-based, GAN-based, and diffusion-based. Regression-based methods (Wu et al., 2025; Balezo et al., 2026) train stably, but their pixel-wise objective drives predictions toward the conditional mean, so the results are over-smoothed and lack high-frequency detail. GAN-based methods (Isola et al., 2017; Liu et al., 2022; Bian et al., 2024; Ma et al., 2026) recover detail but are difficult to stabilize and often fail to converge. Diffusion-based methods (Klöckner et al., 2025; Bin Saleem et al., 2026; Oh et al., 2026)generate higher-quality images, but they operate in a compressed latent space that discards fine histological detail, and their multi-step sampling makes inference slow. Among the three, the diffusion family is the most promising yet the most constrained by the space it generates in.

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.



(a) Comparison of PixStain with the VAE-based latent method

(b) Query similarity maps in VAE, FM, and PFM spaces

Figure 1: Generating in pixel space with pathology foundation model features outperforms latent-space generation. **Left:** prior virtual staining methods generate in a latent space, whereas we generate in pixel space conditioned on pathology foundation model features. **Right:** query similarity maps obtained by using highly expressing regions of different stained images as queries.

This raises a question: is it the latent space or the diffusion formulation itself that limits current models? Our evidence points to the former. As Figure 1 (right) shows, query similarity maps computed in the VAE latent space carry no visible task semantics, and those from a foundation model trained on natural images (Oquab et al., 2023) are only weakly structured, whereas pathology foundation model (PFM) features clearly separate regions of high marker expression. What the latent space discards is the pathology-specific semantics that virtual staining needs to recover, and PFM features supply it in pixel space. This motivates generating in pixel space under PFM conditioning rather than in a VAE latent space, as illustrated in Figure 1 (left). Recent diffusion-based virtual staining models have begun to incorporate PFM features (Bin Saleem et al., 2026), but only as a global conditioning signal. They also leave the feature space unconstrained, supervising the output alone. We therefore propose PixStain, a pixel-space diffusion model built on cross-view representation alignment, which exploits multi-level PFM features to generate high-quality virtual staining results in a single step.

Specifically, we first extract multi-level features from the input H&E image with a PFM and feed them into a pixel-space Vision Transformer, which is trained with flow matching in a single stage to generate the target marker image in a single step. We then introduce cross-view representation alignment, in which the student receives a disjointly corrupted view and the teacher receives the cleaner one. In the student's view, the target marker image is more heavily noised in selected regions, while the conditioning H&E image is masked in spatially disjoint regions. In the teacher's view, the marker is noised only lightly and the H&E image is left intact. The student is trained to match the teacher's high-level features from its own low-level ones. Because information suppressed in one input remains visible in the other, the alignment forces the model to integrate both inputs, yielding stronger representations for virtual staining. We compare our method with regression-based, GAN-based, and diffusion-based models on three paired datasets covering H&E-to-IHC (HER2match), H&E-to-mIHC (HEMIT), and H&E-to-mIF (ORION-CRC), using pixel-level, perceptual, cell classification, and cell segmentation metrics. Our method achieves state-of-the-art performance on all datasets and metric groups, with the sole exception of pixel-level metrics on HEMIT, where it ranks second.

Our contributions are summarized as follows:

- We propose a pixel-space diffusion model for virtual staining that is conditioned on multi-level PFM features and generates high-quality marker images in a single step with single-stage training.
- We propose cross-view representation alignment, which pairs a disjointly corrupted student view with a cleaner teacher view and aligns their low- and high-level representations, consistently improving generation quality.
- We evaluate our method against three families of virtual staining models on three datasets spanning single and multiplex virtual staining, using four groups of metrics. Our method

achieves state-of-the-art results on HER2match and ORION-CRC and the best cell classification performance on HEMIT.

2 Related Work

2.1 Virtual staining

Virtual staining translates histology images between staining modalities using regression-based, GAN-based, or diffusion-based methods. Regression approaches directly predict target stains from H&E: ROSIE (Wu et al., 2025) uses ConvNeXt, while MIPHEI (Balezo et al., 2026) adapts a pathology foundation encoder with LoRA for multiplex immunofluorescence prediction. Their pixel-wise objectives support stable training but favor conditional-mean predictions, often smoothing out high-frequency histological detail.

GAN-based methods use adversarial training to recover fine detail. Building on Pix2Pix (Isola et al., 2017), PyramidPix2Pix (Liu et al., 2022) adds multiscale reconstruction, Pix2PixDualBranch (Bian et al., 2024) combines local CNN features with Swin Transformer context, and MCS-stain (Hu et al., 2025) introduces cell-level semantic guidance. HistoPlexer (Andani et al., 2025) addresses intersection variation through pixel- and embedding-level losses, while DGR (Ma et al., 2026) handles imperfect alignment by decoupling generation from registration.

Diffusion models train more stably and generate higher-quality images, and have accordingly been adopted for virtual staining. BBDM (Klöckner et al., 2025) applies Brownian bridge diffusion to translate directly from source to target. HistDiT (Bin Saleem et al., 2026) conditions a latent diffusion Transformer on H&E spatial latents through cross-attention and on global PFM embeddings through adaptive layer normalization for structure-aware H&E-to-IHC translation. DiffusionFT (Oh et al., 2026) generates multiplex staining channels with one-hot marker conditioning but relies on fine-tuning a pretrained latent diffusion model and VAE decoding, inheriting representations not tailored to virtual staining. In contrast, our method conditions on multi-level features from a PFM rather than on a single global embedding, requiring only single-stage training and single-step inference.

2.2 Pixel-space diffusion model

Latent Diffusion Models (LDM) (Rombach et al., 2022) have become the dominant paradigm for image generation, following a two-stage compress-then-diffuse strategy in which a pretrained VAE maps images into a low-dimensional latent space where the diffusion process is carried out. DiT (Peebles & Xie, 2023) replaces the U-Net backbone with a ViT and achieves state-of-the-art image generation. Because these models operate in the latent space of a VAE, high-frequency textures, sharp boundaries, and fine local details are discarded in the compression step, and a VAE pretrained on natural images faces a further domain gap on pathology images. JiT (Li & He, 2026) shows that a plain ViT with x-prediction suffices to train a diffusion model directly in pixel space. We extend this pixel-space formulation from image generation to image translation and, by conditioning on a pathology foundation model, obtain a single-step pixel-space model for virtual staining.

2.3 Representation Learning in Generative Models

Representation learning has recently gained attention in diffusion models. REPA (Yu et al., 2025) aligns denoising features with external representations to improve efficiency and generation quality. SRA (Jiang et al., 2026) instead aligns shallow, noisier features with deeper, cleaner ones without an external encoder, while Self-Flow (Chefer et al., 2026) applies dual-timestep noising across tokens for representation learning. We extend alignment to cross-view conditioning through spatially disjoint target image noising and H&E masking. Aligning student features with those from a cleaner teacher view encourages the model to integrate complementary information from both inputs.

3 Method

We introduce PixStain, a pixel-space diffusion model with cross-view representation alignment for both single and multiplex virtual staining from H&E in a single step under single-stage training. As illustrated in Figure 2, the model consists of two key components: a pixel-space diffusion model conditioned on multi-level PFM features and trained with flow matching under x-prediction, and a

cross-view representation alignment strategy that aligns the low-level features of the student viewing the disjointly corrupted input with the high-level features of the teacher viewing the cleaner one.

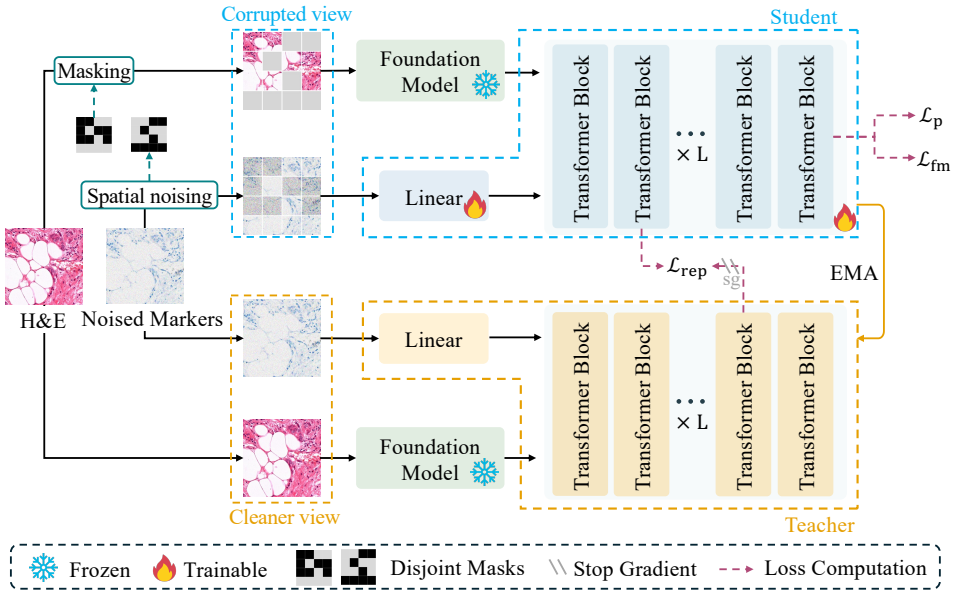


Figure 2: Overview of PixStain, a pixel-space diffusion model conditioned on multi-level PFM features and trained with flow matching under x-prediction. A disjointly corrupted student view is aligned to a cleaner teacher view.

3.1 Pathology-Conditioned Pixel-Space Diffusion

Flow Matching. Flow matching (Liu et al., 2023; Lipman et al., 2023) learns a continuous, time-dependent velocity field that transports a simple source distribution to the target distribution. Let $t \in [0, 1]$ denote a normalized time variable, where $t = 0$ corresponds to the source distribution and $t = 1$ to the target distribution. The source distribution is a standard Gaussian distribution, $p_0(x) = \mathcal{N}(x; \mathbf{0}, \mathbf{I})$, and the target distribution $p_1(x)$ is the target image distribution. Connecting a source sample x_0 and a target sample x_1 along a linear interpolation path gives

$$x_t = (1 - t)x_0 + tx_1 \tag{1}$$

The flow velocity is the time derivative of x_t; for the linear interpolation path above it reduces to the constant $v_t = dx_t/dt = x_1 - x_0$. Flow matching trains a neural network v_θ to approximate this velocity field, so that transporting x_t from $t = 0$ to $t = 1$ carries the source distribution onto the target distribution $p_1(x)$. In our setting the velocity additionally depends on the conditioning H&E image c, and the training objective is

$$\mathcal{L}_{\text{fm}} = \mathbb{E}_{x_0, x_1, t}\left[\left\|v_\theta(x_t, t \mid c) - (x_1 - x_0)\right\|^2\right] \tag{2}$$

Target Marker Image Patchification. Following JiT (Li & He, 2026), we partition the noised target marker image $x_t \in \mathbb{R}^{H \times W \times C}$ into non-overlapping $p \times p$ patches, where the target is a single-marker IHC image ($C = 3$) or a multi-marker mIF or mIHC image ($C \geq 3$). Flattening each patch yields $N = (H/p) \times (W/p)$ tokens of dimension p^2C, which a linear layer then projects into a D_x-dimensional embedding space:

$$T_x = \text{Linear}\left(\text{Patchify}(x_t)\right) \in \mathbb{R}^{N \times D_x} \tag{3}$$

where T_x denotes the sequence of noised target marker tokens.

Multi-level PFM Feature Conditioning. Prior latent diffusion methods draw their H&E conditioning from either a VAE pretrained on natural images, which carries no pathology knowledge (Oh et al., 2026), or the global phenotype embedding of a PFM (Bin Saleem et al., 2026). Both stop

short of the semantically rich patch embeddings that such models produce off the shelf. As Figure 1 shows, the high-level features of the PFM cluster regions of high marker expression, while its low-level features retain the texture detail lost in the high-level semantics. We therefore condition the diffusion model on both.

Specifically, given an H&E image c, let $F_c^k \in \mathbb{R}^{N \times D_c}$ denote the patch tokens from the k-th layer of the pathology foundation model, where D_c is the feature dimension. We uniformly sample L layers $\ell_1, \ldots, \ell_L$ that span low to high levels and concatenate their tokens along the feature dimension:

$$T_c = \text{Concat}\left(F_c^{\ell_1}, F_c^{\ell_2}, \ldots, F_c^{\ell_L}\right) \in \mathbb{R}^{N \times (LD_c)} \tag{4}$$

The resulting conditioning tokens T_c are concatenated with the noised target marker tokens T_x position-wise and fed into the diffusion Transformer.

Target Marker Pixel-space Prediction. Given the target marker image tokens T_x and the multi-level H&E features T_c, we concatenate them position-wise and apply a linear projection to obtain the fused features:

$$Z = \text{Linear}\left(\text{Concat}(T_x, T_c)\right) \tag{5}$$

A stack of Transformer blocks (Peebles & Xie, 2023) processes Z, and a linear output head projects each token back to a $p \times p \times C$ patch, which we reassemble into the clean target marker image $\hat{x}_1$. Following JiT, we train the model with x-prediction and obtain the velocity analytically through the linear interpolation path of Eq. equation 1. We convert this x-prediction into a v-prediction via Eq. equation 6 and supervise it with the flow matching loss in Eq. equation 2:

$$\hat{v}_\theta(x_t, t \mid c) = \frac{\hat{x}_1 - x_t}{1 - t} \tag{6}$$

3.2 Cross-View Representation Alignment

Self-Flow. Strong semantic representations improve the generation quality of diffusion models. Self-Flow (Chefer et al., 2026) introduces a self-supervised flow matching paradigm that integrates representation learning within the generative framework: the model denoises the input and reconstructs the teacher's features from its mixed-noise view at the same time. It samples two timesteps $t, s \sim \pi(t)$ with $t > s$, and a mask

$$M_1^i = \mathbf{1}[\gamma_i < R_{M_1}], \qquad \gamma_i \overset{\text{iid}}{\sim} \text{Uniform}(0,1), \quad i = 1, \ldots, N. \tag{7}$$

Here, R_{M_1} denotes the masking ratio. The mask M_1 corrupts the target image x_1, and the corrupted positions receive the timestep s while the rest receive t, yielding the dual-timestep vector $\tau \in \mathbb{R}^N$:

$$\tau^i = \begin{cases} s, & \text{if } M_1^i = 1, \\ t, & \text{otherwise.} \end{cases} \tag{8}$$

$$\mathbf{x}_\tau = \text{diag}(\mathbf{1} - \tau)\mathbf{x}_0 + \text{diag}(\tau)\mathbf{x}_1 \tag{9}$$

The heterogeneously noised image $\mathbf{x}_\tau$ is the input to a student model f_θ, while the teacher model $f_{\theta'}$ observes the cleaner $\mathbf{x}_{\tau_{\max}}$ noised by $\tau_{\max} = t$. The representation alignment loss is defined using the cosine distance between the student and teacher features::

$$\mathcal{L}_{\text{rep}} = \mathbb{E}_{x_0, x_1, \tau}\left[1 - \cos\left(f_\theta^{(l)}(\mathbf{x}_\tau, \tau),\, f_{\theta'}^{(k)}(\mathbf{x}_{\tau_{\max}}, \tau_{\max})\right)\right] \tag{10}$$

where l and k are the layers of the student and teacher models, respectively, with $l < k$.

Disjoint Masking. Self-Flow considers noising only the target image. In virtual staining, however, noising the target marker alone leaves the conditioning H&E image intact, so the student can match the teacher using that image alone; the alignment task is therefore too easy, and we mask the conditioning image as well. We therefore corrupt the conditioning image with a mask M_2 disjoint from M_1, so that the student's view and the teacher's view differ in both inputs rather than in the target alone. Since the information masked in the two inputs does not overlap, the model must integrate both inputs to accomplish the alignment, which forces it to learn stronger representations.

Specifically, given the mask M_1 used to noise the marker image, we reuse the same γ_i to construct a mask M_2 that corrupts the H&E image c:

$$M_2^i = \mathbf{1}[\gamma_i > 1 - R_{M_2}]. \tag{11}$$

Here, R_{M_2} denotes the masking ratio for the H&E image. Since both masks are thresholded on the same per-token random variables γ_i from opposite ends, they are strictly disjoint, i.e., $M_1 \odot M_2 = \mathbf{0}$, whenever $R_{M_1} + R_{M_2} \leq 1$. Applying M_2 to the conditioning image c yields the masked condition $\tilde{c} = c \odot (\mathbf{1} - M_2)$. The student then takes $(\mathbf{x}_\tau, \tau \mid \tilde{c})$ and the teacher takes $(\mathbf{x}_{\tau_{\max}}, \tau_{\max} \mid c)$.

3.3 Training Objective

In addition to the flow matching loss $\mathcal{L}_{\text{fm}}$ and the feature alignment loss $\mathcal{L}_{\text{rep}}$ introduced above, we employ a perceptual loss to further refine the generated single-marker IHC images. Since our model operates in pixel space, the perceptual loss is applied to the generated image directly, without a VAE decoder. Concretely, we adopt the LPIPS loss:

$$\mathcal{L}_{\text{p}} = \text{LPIPS}\left(\hat{x}_1, x_1\right) \tag{12}$$

where $\hat{x}_1$ is the prediction of student model and x_1 is the ground-truth marker image. The overall training objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{fm}} + \lambda_{\text{rep}}\, \mathcal{L}_{\text{rep}} + \lambda_{\text{p}}\, \mathcal{L}_{\text{p}} \tag{13}$$

where λ_{rep} and λ_{p} balance the three terms.

3.4 Inference

At inference, we sample $x_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ from the source distribution and transport it to $t = 1$ under the conditioning H&E image c, solved with the learned velocity field:

$$\frac{dx_t}{dt} = \hat{v}_\theta(x_t, t \mid c), \tag{14}$$

We discretize $[0, 1]$ into S uniform intervals of size $\Delta = 1/S$ and integrate Eq. equation 14 with the Euler method:

$$x_{t+\Delta} = x_t + \Delta\, \hat{v}_\theta(x_t, t \mid c), \qquad t \in \{0, \Delta, \ldots, 1 - \Delta\}. \tag{15}$$

At $S = 1$ this reduces to a single forward pass, $\hat{x}_1 = x_0 + \hat{v}_\theta(x_0, 0 \mid c)$. In practice, a single step already yields the best results, and quality degrades as S grows (Table 6).

4 Experiments

4.1 Datasets

We evaluate our method on three paired, same-section datasets spanning single and multiplex virtual staining. We do not adopt the widely used MIST (Li et al., 2023) and BCI (Liu et al., 2022) datasets, because their images are taken from adjacent serial sections with limited cellular correspondence, which precludes trustworthy evaluation. Instead, same-section datasets provide stronger cellular correspondence, reducing supervision noise and supporting more reliable evaluation.

HER2match (Klöckner et al., 2025) provides same-section H&E and HER2-IHC images from 17 breast cancer cases. It contains 21,172 paired 1024×1024 tiles at $40\times$ magnification, divided into 11,610/3,582/5,980 training, validation, and test pairs. **HEMIT** (Bian et al., 2024) contains 5,292 cellularly aligned 1024×1024 H&E-to-mIHC patch pairs at $40\times$ magnification from eight colorectal cancer patients. Each H&E patch is paired with a three-channel target comprising DAPI, CD3, and panCK. We use the official patient-disjoint split of 3,717/630/945 training, validation, and test pairs. For both HER2match and HEMIT, we downsample all tiles from $40\times$ to $20\times$ magnification for our experiments. For high-plex evaluation, we use the processed **ORION-CRC** dataset released with MIPHEI-ViT (Lin et al., 2023; Balezo et al., 2026). It contains 41 same-section colorectal cancer WSIs with 16 protein-marker channels. Following the MIPHEI protocol, 37/2/2 WSIs are assigned to training, validation, and testing, yielding 295,096/12,402/10,952 paired 333×333 tiles at $20\times$ magnification. For cell-level classification evaluation on HEMIT and ORION-CRC, we use the nuclei masks and single-cell phenotype labels released by MIPHEI (Balezo et al., 2026).

Evaluation Metrics. We evaluate our model with four categories of metrics. (1) Pixel-level metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and the Pearson Correlation Coefficient (PCC). (2) Perceptual metrics: Learned Perceptual Image Patch Similarity (LPIPS) and Deep Image Structure and Texture Similarity (DISTS), which apply only to RGB images and are therefore reported only for single-marker IHC results. (3) Cell classification metrics: Area Under the Receiver Operating Characteristic curve (AUROC), Area Under the Precision-Recall Curve (AUPRC), balanced accuracy (BAcc), and F1 score. (4) Cell segmentation metrics: Dice score and Aggregated Jaccard Index (AJI).

4.2 Implementation Details

We employ the pretrained H-Optimus-0 (Saillard et al., 2024) model as the H&E foundation encoder and keep all its parameters frozen throughout training, taking its patch tokens from layers 10, 20, 30, and 40. Our diffusion Transformer consists of 12 blocks (Peebles & Xie, 2023) with a patch size of 14, matching the patch grid of the foundation encoder. All remaining trainable components are initialized from scratch without loading pretrained weights. During training, timesteps are drawn from a logit-normal distribution with a time shift of 1.2, and data augmentation follows the MIPHEI protocol (Balezo et al., 2026). We maintain an exponential moving average of the student weights with a decay of 0.9996 and use it as the teacher $f_{\theta'}$; the EMA weights are also used for evaluation. We optimize the model using AdamW with a learning rate of 5×10^{-5}, training for 40 epochs in total with the learning rate linearly warmed up over the first five epochs and then held constant. The masking ratios and loss weights are set to $R_{M_1} = R_{M_2} = 0.25$ and $\lambda_{\text{rep}} = \lambda_{\text{p}} = 0.1$; the student and teacher layers are set to $l = 4$ and $k = 9$. The batch size is 4, and all experiments are conducted on a single NVIDIA GeForce RTX 3090 GPU.

4.3 Comparison with SOTA Methods

To comprehensively evaluate our method, we compare it against three categories of approaches. Regression-based methods include ROSIE (Wu et al., 2025) and MIPHEI (Balezo et al., 2026); GAN-based methods include Pix2Pix (Isola et al., 2017), PyPix2Pix (Liu et al., 2022), TDKstain (Peng et al., 2024), Pix2PixDB (Bian et al., 2024), and MCS-stain (Hu et al., 2025); diffusion-based methods include BBDM (Klöckner et al., 2025) and DiffusionFT (Oh et al., 2026). We evaluate them on the single-marker H&E-to-IHC dataset (HER2match) and on the multi-marker H&E-to-mIHC (HEMIT) and H&E-to-mIF (ORION-CRC) datasets. On HER2match, we report pixel-level, perceptual, and cell segmentation metrics; on HEMIT and ORION-CRC, we report pixel-level and cell classification metrics.

Table 1: Comparison with SOTA methods on HER2match dataset. The **best** and second best performances are highlighted.

Method	Pixel-level Metrics			Perceptual Metrics		Cell Segmentation	
	PSNR↑	SSIM↑	PCC↑	LPIPS↓	DISTS↓	Dice↑	AJI↑
Regression-based							
ROSIE (Wu et al., 2025)	17.44	0.424	0.431	0.661	0.440	0.073	0.027
MIPHEI (Balezo et al., 2026)	18.98	0.458	0.523	0.498	0.352	0.430	0.261
GAN-based							
Pix2Pix (Isola et al., 2017)	18.91	0.388	0.523	0.318	0.261	0.371	0.221
PyPix2Pix (Liu et al., 2022)	17.89	0.389	0.479	0.329	0.237	0.460	0.298
TDKstain (Peng et al., 2024)	17.03	0.349	0.433	0.305	0.228	0.418	0.264
Pix2PixDB (Bian et al., 2024)	18.64	0.417	0.557	**0.276**	0.227	0.553	0.375
MCS-stain (Hu et al., 2025)	18.79	0.456	0.522	0.373	0.270	0.517	0.347
Diffusion-based							
BBDM (Klöckner et al., 2025)	15.29	0.341	0.364	0.354	0.264	0.256	0.133
DiffusionFT (Oh et al., 2026)	18.44	0.467	0.516	0.479	0.338	0.499	0.314
Ours	**20.31**	**0.482**	**0.634**	**0.276**	**0.202**	**0.576**	**0.399**

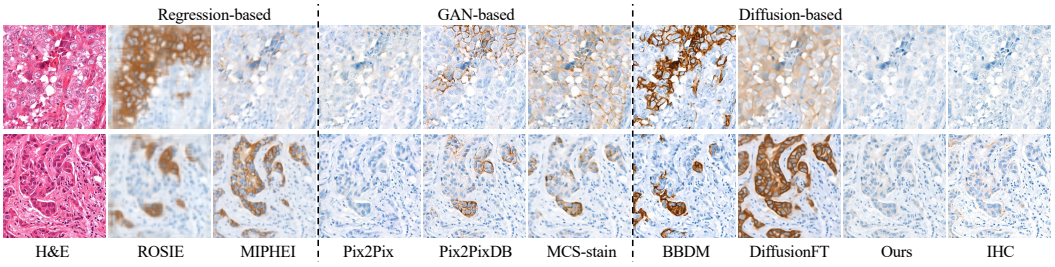


Figure 3: Visual comparison of different methods on the HER2match dataset.

H&E-to-IHC dataset (HER2match). Table 1 reports the comparison on the HER2match dataset. Our method achieves the best performance across all three metric categories: it attains a PSNR of 20.31, an SSIM of 0.482, and a PCC of 0.634, surpassing the runner-up by +1.33, +0.015,

Table 2: Comparison with SOTA methods on HEMIT and ORION-CRC datasets. The **best** and second best performances are highlighted.

Method	HEMIT							ORION-CRC						
	Pixel-level Metrics			Cell Classification Metrics				Pixel-level Metrics			Cell Classification Metrics			
	PSNR	SSIM	PCC	AUROC	AUPRC	BAcc	F1	PSNR	SSIM	PCC	AUROC	AUPRC	BAcc	F1
Regression-based														
ROSIE (Wu et al., 2025)	25.68	0.847	0.630	0.896	0.768	0.831	0.705	30.13	0.941	0.313	0.806	0.384	0.731	0.379
MIPHEI (Balezo et al., 2026)	**30.11**	**0.932**	**0.846**	0.925	0.830	0.856	0.716	31.70	0.950	0.452	0.876	0.502	0.784	0.469
GAN-based														
Pix2Pix (Isola et al., 2017)	25.93	0.867	0.648	0.770	0.555	0.716	0.571	26.16	0.884	0.095	0.683	0.265	0.634	0.292
PyPix2Pix (Liu et al., 2022)	26.15	0.907	0.706	0.890	0.767	0.805	0.654	27.40	0.925	0.266	0.798	0.362	0.722	0.372
Pix2PixDB (Bian et al., 2024)	27.58	0.914	0.768	0.891	0.764	0.827	0.692	27.59	0.895	0.109	0.701	0.275	0.643	0.299
Diffusion-based														
BBDM (Klöckner et al., 2025)	23.64	0.816	0.381	0.757	0.544	0.706	0.554	-	-	-	-	-	-	-
DiffusionFT (Oh et al., 2026)	27.97	0.902	0.743	0.879	0.723	0.810	0.659	30.56	0.946	0.294	0.788	0.376	0.720	0.370
Ours	29.67	0.927	0.841	**0.933**	**0.841**	**0.882**	**0.772**	**31.88**	**0.951**	**0.467**	**0.886**	**0.516**	**0.800**	**0.477**

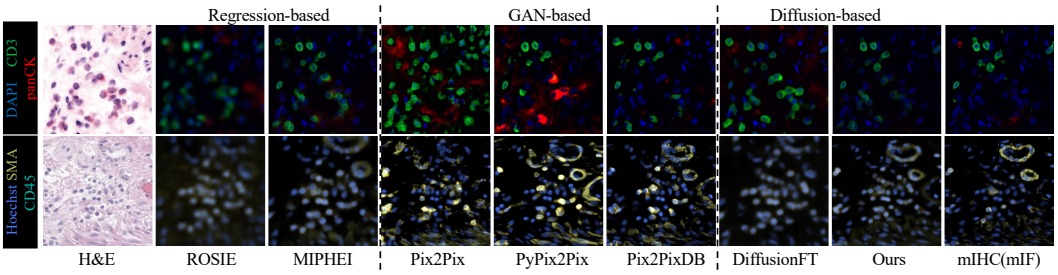


Figure 4: Visual comparison of different methods on the HEMIT and ORION-CRC datasets.

and +0.077, respectively, while its LPIPS of 0.276 matches the best competing method and its DISTS of 0.202 improves upon the runner-up by 0.025. The three families fail in complementary ways. Regression-based methods obtain competitive pixel-level scores but markedly inferior perceptual quality: trained with pixel-level losses, they favor low-frequency content and produce over-smoothed outputs. GAN-based methods exhibit the opposite trend, achieving strong perceptual scores while lagging behind on several pixel-level metrics. The diffusion-based DiffusionFT remains mediocre in both respects. Only our method ranks first across both dimensions. The qualitative comparison in Figure 3 corroborates these observations: regression-based methods produce visibly over-smoothed outputs, whereas our method preserves both sharp details and faithful staining patterns. Visualization of cell segmentation results is shown in Appendix B.2.

H&E-to-mIHC and mIF datasets (HEMIT and ORION-CRC). Table 2 reports the comparison on the HEMIT and ORION-CRC datasets. On HEMIT, our method trails MIPHEI (Balezo et al., 2026) by 0.44 on PSNR and 0.005 on SSIM and PCC, yet it achieves the best scores on all four cell classification metrics, improving over the runner-up by 0.025 on average and by 0.056 on F1. Despite a marginal gap at the pixel level, our method delivers substantial gains at the cell level, which is the quantity of interest for pathology-oriented analysis. On ORION-CRC, our method ranks first on both pixel-level and cell classification metrics: it attains a PSNR of 31.88, an SSIM of 0.951, and a PCC of 0.467, surpassing the runner-up by 0.18, 0.001, and 0.015, respectively, and improving the cell classification metrics by 0.012 on average. Notably, on both mIHC and mIF, the two foundation-model-based methods, ours and MIPHEI, occupy the top two positions, highlighting the strength of pathology foundation models in modeling the complex TME. Additional per-marker results are provided in Appendix B.1, and visual comparisons in Figure 4.

Table 3: Ablation Studies of H&E Features, Representation Alignment, and Perceptual Loss.

Method	HER2match							HEMIT						
	Pixel-level Metrics			Perceptual Metrics		Cell Segmentation		Pixel-level Metrics			Cell Classification Metrics			
	PSNR↑	SSIM↑	PCC↑	LPIPS↓	DISTS↓	Dice↑	AJI↑	PSNR↑	SSIM↑	PCC↑	AUROC↑	AUPRC↑	BAcc↑	F1↑
+H&E Features														
Linear	**19.12**	0.458	0.519	0.477	0.350	0.468	0.286	28.31	0.915	0.807	0.900	0.785	0.801	0.668
Multi-layer PFM Feature	**19.12**	**0.469**	**0.561**	**0.423**	**0.311**	**0.520**	**0.339**	**29.47**	**0.925**	**0.835**	**0.926**	**0.830**	**0.874**	**0.758**
+Representation alignment														
Self-Flow (Chefer et al., 2026)	19.39	0.472	0.563	**0.442**	0.323	0.515	0.333	29.44	**0.927**	0.839	0.928	0.833	0.867	0.742
cross-view representation alignment	**19.86**	**0.480**	**0.602**	0.444	**0.319**	**0.518**	**0.335**	**29.67**	**0.927**	**0.841**	**0.933**	**0.841**	**0.882**	**0.772**
+Perceptual loss														
LPIPS	20.31	0.482	0.634	0.276	0.202	0.576	0.399	-	-	-	-	-	-	-

4.4 Ablation Studies

In this section, we conduct ablation studies on HER2match and HEMIT to evaluate the contribution of the H&E feature, representation alignment, and the perceptual loss. We then analyze the H&E feature conditioning and the H&E mask strategy, and finally examine the effect of the number of inference steps. The results are presented in Table 3. Replacing the linear H&E feature with the multi-layer PFM feature improves performance on both datasets. The gain is most pronounced on HEMIT's cell-classification F1, which rises by 0.09, and on HER2match the perceptual metrics improve as well, with LPIPS and DISTS dropping by 0.054 and 0.039, respectively. Among the pixel-level metrics, PCC shows the largest gain, rising by 0.042 and 0.028 on the two datasets. This indicates that the multi-layer PFM feature captures not only the semantic information of H&E but also its texture information. Next, we incorporate a feature alignment module. Directly adopting Self-Flow (Chefer et al., 2026) improves some metrics and degrades others on both datasets. Compared with Self-Flow, our cross-view alignment improves most reported metrics on both datasets. Finally, We apply perceptual loss only to HER2match, where it substantially improves the perceptual metrics, with LPIPS and DISTS dropping by 0.168 and 0.117, respectively.

H&E feature conditioning. Table 4 compares different H&E feature conditioning strategies. We first compare DINOv2 (Oquab et al., 2023), a foundation model trained on natural images with a comparable parameter count, against H-Optimus-0 (Saillard et al., 2024), which carries pathology-specific knowledge. H-Optimus-0 outperforms DINOv2 on every metric, indicating that pathology-specific pretraining is important for virtual staining. We then examine how the number of feature layers affects performance: Using more feature layers generally improves performance, with four layers achieving the best results across all reported metrics. Thus we adopt four layers to balance efficiency and performance. Finally, we compare feature fusion strategies and find that concatenation outperforms mean fusion.

Table 4: Comparison of different H&E feature conditioning strategies on HER2match dataset.

Settings	Pixel-level Metrics			Perceptual Metrics		Cell Segmentation	
	PSNR↑	SSIM↑	R↑	LPIPS↓	DISTS↓	Dice↑	AJI↑
Foundation model							
DINOv2	19.73	0.471	0.589	0.286	0.209	0.553	0.375
H-Optimus-0	**20.31**	**0.482**	**0.634**	**0.276**	**0.202**	**0.576**	**0.399**
Number of layers							
1	19.67	0.450	0.573	0.297	0.214	0.518	0.338
2	19.58	0.457	0.576	0.291	0.212	0.531	0.352
3	19.89	0.460	0.599	0.281	0.204	0.540	0.363
4	**20.31**	**0.482**	**0.634**	**0.276**	**0.202**	**0.576**	**0.399**
Feature fusion							
Mean	19.94	0.468	0.609	0.283	0.203	0.551	0.372
Concat	**20.31**	**0.482**	**0.634**	**0.276**	**0.202**	**0.576**	**0.399**

Table 5: Comparison of Different H&E Masking Strategies on HER2match dataset.

H&E Masking	Pixel-level Metrics			Perceptual Metrics		Cell Segmentation	
	PSNR↑	SSIM↑	PCC↑	LPIPS↓	DISTS↓	Dice↑	AJI↑
Unmasked	20.03	0.465	0.605	0.276	0.206	0.550	0.373
Same as M_1	19.70	0.464	0.589	0.293	0.210	0.550	0.370
Disjoint (M_2)	**20.31**	**0.482**	**0.634**	**0.276**	**0.202**	**0.576**	**0.399**

Table 6: Comparison of Different Inference Steps on HER2match

Steps	Pixel-level Metrics			Perceptual Metrics		Cell Segmentation	
	PSNR↑	SSIM↑	PCC↑	LPIPS↓	DISTS↓	Dice↑	AJI↑
1	**20.31**	**0.482**	**0.634**	**0.276**	0.202	**0.576**	**0.399**
2	20.09	0.477	0.619	0.282	0.198	0.575	0.396
3	19.86	0.468	0.601	0.294	**0.197**	0.571	0.392
5	19.49	0.451	0.573	0.316	0.208	0.560	0.383

Disjoint Masking. Table 5 compares three masking strategies for the conditioning image, all trained with the alignment loss: leaving the H&E image unmasked, masking it with the same mask M_1 used to noise the target marker, and masking it with a disjoint mask M_2. The disjoint strategy improves every metric group. Sharing M_1 instead degrades PSNR, SSIM, and PCC below even the unmasked baseline, indicating that corrupting the conditioning image is not enough: what matters is that the two inputs stay informative in complementary regions.

Inference Steps. Finally, Table 6 reports the effect of the number of inference steps. A single step already attains the best performance, and quality degrades monotonically with more steps on all metrics except DISTS, which is lowest at 3 steps but varies by at most 0.011 throughout.

5 Conclusion

We presented PixStain, a pixel-space diffusion model for virtual staining. PixStain is conditioned on multi-level PFM features and trained with flow matching under x-prediction in a single stage, and its cross-view representation alignment generates high-quality results in a single step, for both single and multiplex targets from H&E. On three publicly available same-section datasets covering H&E-to-IHC, H&E-to-mIHC, and H&E-to-mIF, PixStain achieves state-of-the-art performance across pixel-level, perceptual, cell classification, and cell segmentation metrics, with the sole exception of pixel-level metrics on HEMIT, where it ranks second.

AI USE STATEMENT

We used generative AI tools for two tasks: assisting with code generation and aiding and polishing writing. We have reviewed all AI-assisted work: generated code was tested for correctness by the authors, and all technical claims, experimental results, and citations were verified against our own experiment logs and the original papers. No research ideas, experimental results, or figures were generated by AI tools. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Faisal R Ali, Guy E Orchard, and Rajeev Mallipeddi. Hematoxylin in history—the heritage of histology. *JAMA Dermatology*, 153(3):328–328, 2017.

Sonali Andani, Boqi Chen, Joanna Ficek-Pascual, Simon Heinke, Ruben Casanova, Bernard Friedrich Hild, Bettina Sobottka, Bernd Bodenmiller, Viktor H Koelzer, et al. Histopathology-based protein multiplex generation using deep learning. *Nature Machine Intelligence*, 7(8):1292–1307, 2025.

Guillaume Balezo, Roger Trullo, Albert Pla Planas, Etienne Decencière, and Thomas Walter. Miphei-vit: Multiplex immunofluorescence prediction from h&e images using vit foundation models. *Computers in Biology and Medicine*, 206:111564, 2026.

John D Bancroft and Marilyn Gamble. *Theory and practice of histological techniques*. Elsevier health sciences, 2008.

Chang Bian, Beth Phillips, Tim Cootes, and Martin Fergie. Hemit: 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.

Aasim Bin Saleem, Amr Ahmed, Ardhendu Behera, Hafeezullah Amin, Iman Yi Liao, Mahmoud Khattab, Pan Jia Wern, and Haslina Makmur. Histdit: A structure-aware latent conditional diffusion model for high-fidelity virtual staining in histopathology. In *International Conference on Pattern Recognition*, pp. 416–431. Springer, 2026.

Hila Chefer, Patrick Esser, Dominik Lorenz, Dustin Podell, Vikash Raja, Vinh Tong, Antonio Torralba, and Robin Rombach. Self-supervised flow matching for scalable multi-modal synthesis. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=HoThWhfxiK.

Yihuang Hu, Zhicheng Du, Weiping Lin, Shurong Yang, Lequan Yu, Guojun Zhang, and Liansheng Wang. Mcs-stain: Boosting ffpe-to-he virtual staining with multiple cell semantics. *IEEE Transactions on Medical Imaging*, 2025.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 5967–5976. Ieee, 2017.

Dengyang Jiang, Mengmeng Wang, Liuzhuozheng Li, Lei Zhang, Haoyu Wang, Wei Wei, Guang Dai, Yanning Zhang, and Jingdong Wang. Representation alignment for diffusion transformers without external components. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ds5w2xth93.

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. In *MICCAI Workshop on Deep Generative Models*, pp. 120–130. Springer, 2025.

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, 2023.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36115–36125, 2026.

Jia-Ren Lin, Yu-An Chen, Daniel Campton, Jeremy Cooper, Shannon Coy, Clarence Yapp, Juliann B Tefft, Erin McCarty, Keith L Ligon, Scott J Rodig, et al. High-plex immunofluorescence imaging and traditional histology of the same tissue section for discovering image-based biomarkers. *Nature cancer*, 4(7):1036–1052, 2023.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Shengjie Liu, Chuang Zhu, Feng Xu, Xinyu Jia, Zhongyue Shi, and Mulan Jin. Bci: Breast cancer immunohistochemical image generation through pyramid pix2pix. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1814–1823. IEEE, 2022.

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

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.

Hyun-Jic Oh, Junsik Kim, Zhiyi Shi, Yichen Wu, Yu-An Chen, Peter K Sorger, Hanspeter Pfister, and Won-Ki Jeong. Virtual multiplex staining for histological images using a marker-wise conditioned diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8168–8176, 2026.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

Marius Pachitariu, Michael Rariden, and Carsen Stringer. Cellpose-sam: superhuman generalization for cellular segmentation. *BioRxiv*, pp. 2025–04, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

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.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Charlie Saillard, Rodolphe Jenatton, Felipe Llinares-López, Zelda Mariet, David Cahané, Eric Durand, and Jean-Philippe Vert. H-optimus-0, 2024. URL https://github.com/bioptimus/releases/tree/main/models/h-optimus/v0.

Eric Wu, Matthew Bieniosek, Zhenqin Wu, Nitya Thakkar, Gregory W Charville, Ahmad Makky, Christian M Schürch, Jeroen R Huyghe, Ulrike Peters, Christopher I Li, et al. Rosie: Ai generation of multiplex immunofluorescence staining from histopathology images. *Nature Communications*, 16(1):7633, 2025.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations*, 2025.

A Evaluation Details

In this section, we provide additional details on the cell segmentation and cell classification evaluation protocols.

Cell Segmentation Evaluation Details. We evaluate cell segmentation on HER2match using Cellpose-SAM (Pachitariu et al., 2025). For each paired sample, we apply Cellpose-SAM to both the real and generated IHC images. The segmentation masks extracted from the real IHC images serve as the ground-truth reference, while those extracted from the generated IHC images are treated as predictions. We compute the Dice similarity coefficient (Dice) and Aggregated Jaccard Index (AJI) between these masks to assess how well the generated images preserve cellular structures. The reference masks are thus obtained automatically from the real images rather than from manual annotations.

Cell Classification Evaluation Details. We evaluate cell classification on HEMIT and ORION-CRC. Following Balezo et al. (2026), we use the released nuclei masks to extract the mean predicted intensity of each marker within each segmented cell. The resulting vector of marker intensities serves as the input to a logistic-regression classifier, with reference cell phenotype labels derived from the measured multiplex images. For each dataset, the classifier is trained on validation-set cells and evaluated on test-set cells. We report the AUPRC, AUROC, BAcc, and F1 score for each cell phenotype.

B More Experiment Results

B.1 Quantitative Results

In this section, we provide detailed per-marker results for HEMIT and ORION-CRC in Tables 7 and 9, respectively. We also report per-phenotype results for the downstream cell classification task on these two datasets in Tables 8 and 10, respectively.

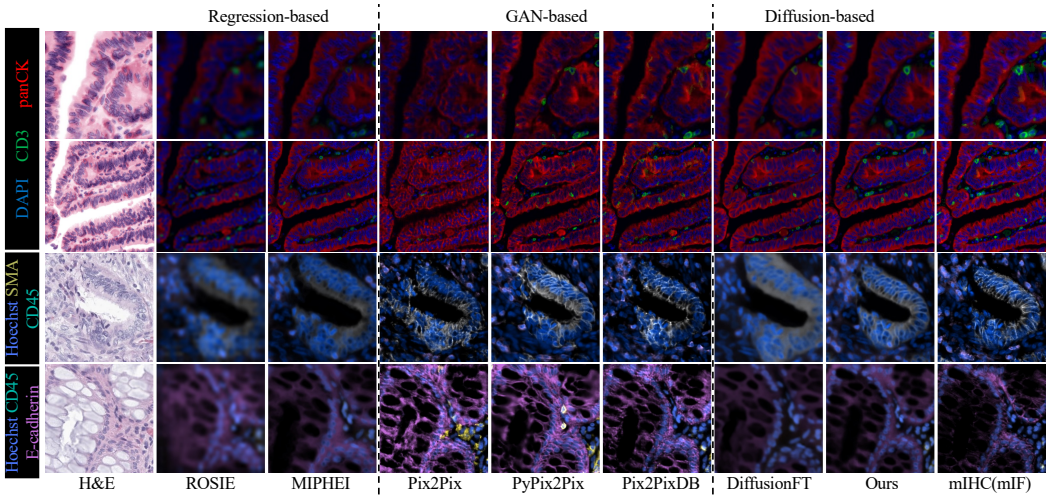


Figure 5: Visual results of our method compared to the state-of-the-art methods on the HEMIT and ORION-CRC datasets.

B.2 Visual Results

In this section, we present additional visual examples of generated images on HEMIT and ORION-CRC in Figure 5, followed by qualitative comparisons of cell segmentation results between our method and state-of-the-art methods on HER2match in Figure 6.

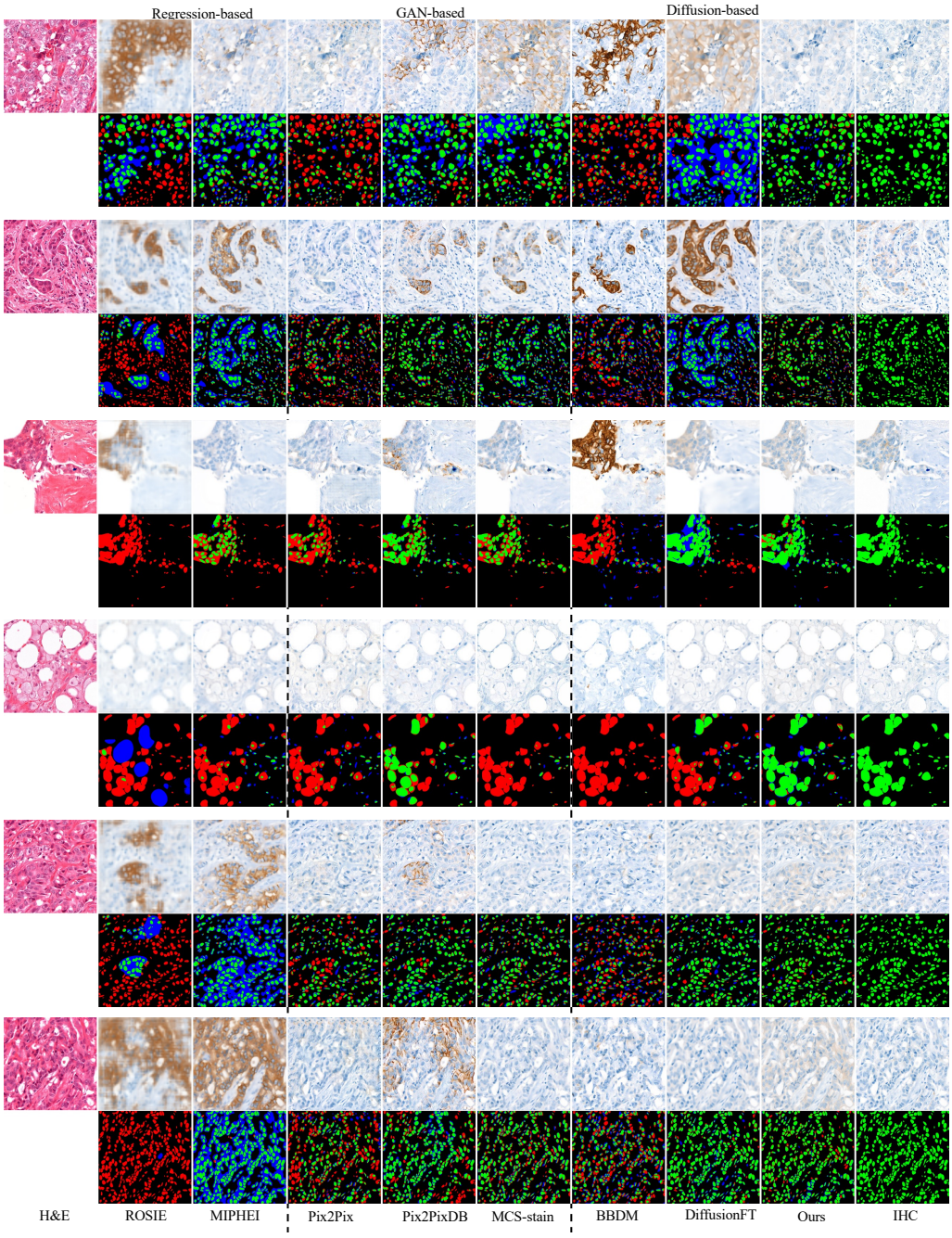


Figure 6: Qualitative comparison of cell segmentation results on the HER2match dataset between our method and state-of-the-art methods. Green: true positives; Red: false negatives; Blue: false positives.

Table 7: HEMIT dataset

Method	Avg.			CD3			DAPI			panCK		
	PSNR	SSIM	R	PSNR	SSIM	R	PSNR	SSIM	R	PSNR	SSIM	R
Regression-based												
ROSIE (Wu et al., 2025)	25.68	0.847	0.630	24.45	0.879	0.530	28.17	0.802	0.689	24.43	0.861	0.672
MIPHEI (Balezo et al., 2026)	**30.11**	**0.932**	**0.846**	25.99	**0.904**	0.713	**36.46**	**0.964**	**0.972**	**27.88**	**0.926**	**0.853**
GAN-based												
Pix2Pix (Isola et al., 2017)	25.93	0.867	0.648	21.35	0.822	0.374	31.73	0.926	0.905	24.70	0.853	0.666
PyPix2Pix (Liu et al., 2022)	26.15	0.907	0.706	23.58	0.886	0.550	29.86	0.947	0.832	25.01	0.889	0.737
Pix2PixDB (Bian et al., 2024)	27.58	0.914	0.768	24.53	0.892	0.568	32.26	0.946	0.960	25.94	0.903	0.776
Diffusion-based												
BBDM (Klöckner et al., 2025)	23.64	0.816	0.381	21.72	0.850	0.153	26.50	0.762	0.530	22.70	0.839	0.463
DiffusionFT (Oh et al., 2026)	27.97	0.902	0.743	23.68	0.866	0.477	33.71	0.933	0.948	26.51	0.908	0.803
Ours	29.67	0.927	0.841	**26.04**	**0.904**	**0.715**	35.43	0.955	0.968	27.52	0.923	0.842

Table 8: HEMIT downstream task

Method	Avg.				CD3				panCK			
	AUROC	AUPRC	BAcc	F1	AUROC	AUPRC	BAcc	F1	AUROC	AUPRC	BAcc	F1
Regression-based												
ROSIE	0.896	0.768	0.831	0.705	0.879	0.657	0.808	0.580	0.912	0.879	0.855	0.830
MIPHEI	0.925	0.830	0.856	0.716	0.885	0.706	0.787	0.519	**0.965**	**0.955**	**0.925**	**0.912**
GAN-based												
Pix2Pix	0.770	0.555	0.716	0.571	0.622	0.227	0.587	0.323	0.918	0.883	0.846	0.819
PyPix2Pix	0.890	0.767	0.805	0.654	0.835	0.625	0.718	0.440	0.945	0.909	0.891	0.869
Pix2PixDB	0.891	0.764	0.827	0.692	0.827	0.601	0.747	0.496	0.955	0.927	0.907	0.888
Diffusion-based												
BBDM	0.757	0.544	0.706	0.554	0.658	0.290	0.607	0.339	0.856	0.797	0.806	0.769
DiffusionFT	0.879	0.723	0.810	0.659	0.813	0.514	0.710	0.423	0.945	0.932	0.910	0.895
Ours	**0.933**	**0.841**	**0.882**	**0.772**	**0.909**	**0.732**	**0.843**	**0.639**	0.957	0.950	0.920	0.906

Table 9: ORION dataset

Method	Metric	Avg.	Hoechst	CD31	CD45	CD68	CD4	FOXP3	CD8a	CD45RO	CD20	PD-L1	CD3e	CD163	E-cadherin	Ki67	Pan-CK	SMA
Regression-based																		
ROSIE	SSIM	0.941	0.772	**0.988**	0.919	0.969	0.946	**0.998**	0.993	0.987	**0.997**	0.989	**0.974**	0.967	0.821	**0.978**	0.874	0.877
	R	0.313	0.861	0.125	0.448	0.134	0.345	0.085	0.174	0.276	0.291	0.034	0.421	0.097	0.413	0.297	0.715	0.291
	PSNR	30.13	22.94	34.64	25.22	29.53	26.79	43.25	36.35	34.55	40.37	32.46	32.12	26.78	18.89	31.33	23.43	23.47
MIPHEI	SSIM	0.950	0.856	**0.988**	0.926	0.971	**0.948**	**0.998**	0.992	**0.988**	**0.997**	0.993	**0.974**	0.981	0.836	0.975	0.888	**0.884**
	R	0.452	0.915	**0.299**	0.643	0.316	0.537	0.155	0.231	0.439	0.487	0.127	0.542	0.285	**0.633**	0.404	0.800	0.418
	PSNR	31.70	25.21	35.15	27.37	**30.38**	27.69	**43.26**	36.55	**35.93**	41.68	37.41	32.72	**33.24**	**20.14**	31.70	24.67	24.09
GAN-based																		
Pix2Pix	SSIM	0.884	0.517	0.974	0.804	0.927	0.875	0.991	0.977	0.958	0.994	0.977	0.932	0.943	0.742	0.939	0.805	0.784
	R	0.095	0.282	0.011	0.156	0.026	0.096	0.001	0.017	0.067	0.139	-0.002	0.057	0.012	0.164	0.050	0.343	0.094
	PSNR	26.16	15.85	31.37	21.32	25.20	23.27	35.52	30.10	28.98	37.85	31.61	26.97	26.66	17.27	26.17	19.76	20.73
PyPix2Pix	SSIM	0.925	0.816	0.972	0.863	0.928	0.915	0.993	0.983	0.972	0.994	0.980	0.956	0.950	0.812	0.966	0.866	0.838
	R	0.266	0.878	0.082	0.407	0.098	0.295	0.050	0.085	0.155	0.220	0.071	0.295	0.104	0.402	0.204	0.677	0.237
	PSNR	27.40	23.27	29.03	22.76	23.86	24.76	35.58	32.09	30.59	37.30	31.02	29.16	27.05	18.60	28.95	22.53	21.84
Pix2PixDB	SSIM	0.895	0.512	0.975	0.834	0.927	0.905	0.995	0.982	0.978	0.995	0.980	0.950	0.958	0.755	0.961	0.798	0.814
	R	0.109	0.294	0.014	0.180	0.031	0.116	0.002	0.015	0.111	0.142	0.016	0.087	0.016	0.195	0.037	0.381	0.109
	PSNR	27.59	15.81	31.26	22.43	25.77	24.75	38.83	32.61	32.44	38.17	32.97	28.88	29.61	17.68	28.86	19.68	21.60
Diffusion-based																		
DiffusionFT	SSIM	0.946	0.829	**0.988**	0.917	**0.972**	0.947	**0.998**	**0.994**	0.987	0.996	0.990	**0.974**	**0.982**	0.825	0.969	0.881	**0.884**
	R	0.294	0.891	0.069	0.432	0.123	0.372	0.054	0.145	0.286	0.237	-0.012	0.381	0.079	0.399	0.318	0.722	0.208
	PSNR	30.56	24.14	34.80	24.96	29.93	26.96	43.19	**36.85**	34.63	40.20	32.39	31.94	32.66	18.80	30.47	23.49	23.45
Ours	SSIM	**0.951**	**0.869**	0.987	**0.928**	0.968	0.947	0.997	0.992	0.987	**0.997**	**0.994**	0.973	0.979	**0.840**	0.976	**0.894**	0.883
	R	**0.467**	**0.924**	0.297	**0.662**	**0.318**	**0.547**	**0.159**	**0.251**	**0.446**	**0.507**	**0.221**	**0.560**	**0.308**	0.588	**0.421**	**0.818**	**0.448**
	PSNR	**31.88**	**25.62**	**35.21**	**27.53**	30.31	**27.80**	43.05	36.80	35.86	**41.76**	**38.96**	**32.82**	33.17	19.89	**32.05**	**24.98**	**24.25**

Table 10: Performance comparison across methods and markers.

Method	Metric	Avg.	CD31	CD45	CD68	CD4	FOXP3	CD8a	CD45RO	CD20	PD-L1	CD3e	CD163	E-cadherin	Ki67	Pan-CK	SMA
Regression-based																	
ROSIE	AUROC	0.806	0.804	0.801	0.762	0.817	0.864	0.835	0.856	0.794	0.517	0.824	0.760	0.890	0.807	0.918	0.838
	AUPRC	0.384	0.145	0.552	0.135	0.458	0.228	0.165	0.395	0.253	0.011	0.498	0.104	0.908	0.306	0.938	0.658
	BAcc	0.731	0.720	0.717	0.704	0.737	0.769	0.740	0.749	0.717	0.495	0.739	0.689	0.825	0.766	0.837	0.764
	F1	0.379	0.173	0.584	0.219	0.487	0.128	0.122	0.377	0.194	0.018	0.503	0.161	0.802	**0.414**	0.821	0.680
MIPHEI	AUROC	0.876	0.895	0.885	0.833	0.881	0.911	0.882	0.921	0.862	0.683	0.888	0.875	**0.960**	0.795	0.969	0.898
	AUPRC	0.502	0.365	0.758	**0.271**	**0.640**	0.319	0.195	0.603	0.336	0.025	**0.671**	0.284	**0.964**	0.316	0.975	0.814
	BAcc	0.784	0.782	0.782	0.747	0.780	0.813	0.778	0.799	0.754	0.610	0.787	0.787	0.908	0.698	0.918	0.814
	F1	0.469	**0.349**	0.686	**0.316**	0.607	**0.216**	0.180	**0.552**	0.285	0.040	0.622	0.259	0.903	0.371	0.917	0.737
GAN-based																	
Pix2Pix	AUROC	0.683	0.652	0.667	0.630	0.698	0.704	0.675	0.707	0.678	0.518	0.695	0.656	0.731	0.753	0.760	0.728
	AUPRC	0.265	0.061	0.396	0.090	0.305	0.078	0.052	0.179	0.120	0.010	0.315	0.067	0.740	0.247	0.784	0.535
	BAcc	0.634	0.620	0.629	0.602	0.651	0.643	0.591	0.623	0.596	0.491	0.649	0.632	0.695	0.709	0.710	0.673
	F1	0.292	0.103	0.479	0.155	0.384	0.068	0.061	0.227	0.107	0.017	0.394	0.125	0.637	0.378	0.659	0.583
PyPix2Pix	AUROC	0.798	0.760	0.793	0.757	0.808	0.854	0.820	0.846	0.775	0.669	0.812	0.762	0.840	0.748	0.875	0.848
	AUPRC	0.362	0.132	0.528	0.141	0.428	0.188	0.119	0.330	0.224	0.017	0.458	0.114	0.875	0.225	0.912	0.733
	BAcc	0.722	0.674	0.718	0.694	0.718	0.760	0.721	0.754	0.707	0.592	0.719	0.690	0.807	0.691	0.826	0.762
	F1	0.372	0.151	0.588	0.225	0.490	0.139	0.124	0.378	0.180	0.032	0.503	0.161	0.778	0.352	0.800	0.677
Pix2PixDB	AUROC	0.701	0.676	0.681	0.632	0.710	0.716	0.706	0.721	0.715	0.534	0.710	0.682	0.755	0.741	0.789	0.741
	AUPRC	0.275	0.071	0.417	0.094	0.320	0.076	0.054	0.202	0.125	0.012	0.331	0.078	0.751	0.231	0.800	0.560
	BAcc	0.643	0.639	0.609	0.602	0.656	0.648	0.607	0.630	0.615	0.513	0.656	0.650	0.705	0.711	0.726	0.682
	F1	0.299	0.108	0.438	0.158	0.391	0.071	0.072	0.247	0.138	0.021	0.402	0.135	0.656	0.374	0.687	0.593
Diffusion-based																	
DiffusionFT	AUROC	0.788	0.728	0.775	0.752	0.810	0.858	0.833	0.800	0.753	0.504	0.815	0.790	0.838	**0.825**	0.880	0.854
	AUPRC	0.376	0.077	0.524	0.148	0.486	0.206	0.148	0.276	0.239	0.010	0.513	0.141	0.881	**0.339**	0.914	0.731
	BAcc	0.720	0.674	0.706	0.688	0.719	0.768	0.743	0.706	0.687	0.477	0.721	0.710	0.818	**0.767**	0.834	0.776
	F1	0.370	0.127	0.568	0.206	0.482	0.152	0.144	0.294	0.154	0.016	0.500	0.179	0.796	0.411	0.821	0.695
Ours	AUROC	**0.886**	**0.905**	**0.892**	**0.844**	**0.885**	**0.919**	**0.893**	**0.927**	**0.870**	**0.733**	**0.893**	**0.880**	0.958	0.813	**0.972**	**0.909**
	AUPRC	**0.516**	**0.394**	**0.762**	0.267	0.632	**0.334**	**0.213**	**0.618**	**0.386**	**0.051**	0.669	**0.305**	**0.964**	**0.339**	**0.978**	**0.836**
	BAcc	**0.800**	**0.791**	**0.798**	**0.762**	**0.794**	**0.833**	**0.799**	**0.820**	**0.769**	**0.650**	**0.801**	**0.788**	**0.913**	0.729	**0.923**	**0.828**
	F1	**0.477**	0.348	**0.705**	**0.316**	**0.612**	0.215	**0.187**	0.551	**0.293**	**0.051**	**0.630**	**0.265**	**0.908**	0.396	**0.922**	**0.757**