STAINFLASH: DEEP-COMPRESSION LINEAR-ATTENTION FLOW MATCHING WITH PATHOLOGICAL REPRESENTATION FUSION FOR H&E-TO-IHC VIRTUAL STAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Immunohistochemical (IHC) staining reveals molecular information essential for accurate pathological diagnosis, yet it requires costly and time-consuming chemical procedures. Virtual staining computationally synthesizes IHC images from hematoxylin-eosin (H&E) images stained with low chemical cost, offering a cost-effective alternative. Despite remarkable progress in virtual staining based on recent diffusion models, achieving high-quality, high-resolution H&E-to-IHC staining with low inference latency and computational cost remains a challenge. This paper presents StainFlash, an efficient framework that leverages a flow-matching network with a deep compression autoencoder and linear attention to overcome the latency and memory constraints of high-resolution virtual staining. To enhance pathological fidelity, we propose a timestep-aware lightweight pathology encoder that adaptively modulates the layout embedding extracted from H&E images, along with a novel **k**nowledge-**i**ntensive **i**mplicit representations **f**usion (KIIF) module that effectively fuses knowledge-intensive implicit representations from multiple pathology foundation models into the flow matching network. Extensive experiments on four datasets stained with various biomarkers demonstrate that StainFlash outperforms existing approaches in high-quality, high-resolution H&E-to-IHC staining, with substantially reduced latency and computational cost. Ablation studies also validate StainFlash's novel designs. Our code and trained models will be released.

1 INTRODUCTION

Histochemical staining is widely used to distinguish various tissues by coloring them, thereby facilitating pathological diagnosis Stack et al. (2014). Hematoxylin-eosin (H&E) stains are routinely used to highlight cellular morphology and delineate tissue structures. However, H&E may not always provide sufficient diagnostic information, particularly protein-specific expression critical for tumor identification and cancer prognosis Bai et al. (2023). Immunohistochemistry (IHC) is widely used to reveal distribution and localization of specific markers, offering essential insights for tumor subtyping, prognosis, and personalized treatment. However, IHC staining requires a much more complex and costly preparation process. In clinical practice, acquiring paired H&E-IHC images often involves repeated serial sectioning and application of different antibodies, a time-consuming and tissue-intensive process.

Virtual staining presents a promising, cost-effective alternative to conventional chemical staining for generating multiple stains. Essentially an image-to-image translation task, it digitally "translates" chemically stained pathology images, converting images from the source domain (e.g., H&E) into new ones resembling those of the target domain (e.g., IHC), while preserving cellular morphology, tissue structure, and pathological semantics. Ideally, a virtually stained image should exhibit high perceptual quality (e.g., clear nuclei, and color and texture distributions closely aligned with the target domain) as well as high pathological fidelity (e.g., textual accuracy in the distribution and size of nuclei, and pathological consistency in lesion localization and cancer severity).

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.

Researchers have employed generative adversarial networks (GANs; Goodfellow et al., 2014) for virtual staining Peng et al. (2024). Despite notable progress, GANs may encounter significant training challenges, such as mode collapse Thanh-Tung & Tran (2020). Recent studies have explored diffusion models Esser et al. (2024) for virtual staining, either in the pixel space Chen et al. (2025c); He et al. (2024); Kataria et al. (2024; 2025); Klöckner et al. (2025); Liu et al. (2025a); Xiong et al. (2025); Zheng et al. (2025) or latent space Dubey et al. (2024); Großkopf et al. (2025); Ho et al. (2025); Jewsbury et al. (2024); Yellapragada et al. (2025a). However, when generating high-resolution outputs (e.g., 1024×1024 pixels or higher), these methods suffer from high inference latency and may even fail to produce results due to memory constraints.

This work presents StainFlash, an efficient, high-resolution virtual staining framework based on flow matching networks Liu et al. (2022b) with a deep compression autoencoder (DC-AE; Chen et al., 2025b) and linear attention Katharopoulos et al. (2020). To address both inference latency and memory constraints, StainFlash adopts a DC-AE with a $32\times$ spatial compression ratio (vs. $8\times$ in conventional autoencoders Esser et al. (2024)), significantly shortening token sequences for more efficient attention processing. Additionally, StainFlash introduces linear-complexity attention into the flow matching network following the paradigm in text-to-image generation models such as SANA Xie et al. (2025). Compared to the softmax attention, StainFlash reduces the attention computation complexity from $\mathcal{O}(N^2)$ to $\mathcal{O}(N)$, directly improving computational speed. Together, these innovations make StainFlash highly scalable for high-resolution virtual staining, overcoming the limitations of latency and memory demand. We further design a lightweight pathological encoder that efficiently extracts layout embeddings from H&E images to guide the flow matching network with learned timestep-dependent weights. To the best of our knowledge, StainFlash is the first work to combine DC-AE and linear attention with flow matching for virtual staining.

Recent pathology foundation models Chen et al. (2024b); Filiot et al. (2023; 2024); kaiko.ai et al. (2024); Kang et al. (2023); Karasikov et al. (2025); Nechaev et al. (2024); Xu et al. (2024); Zimmermann et al. (2024), trained on massive pathology data with a self-supervised paradigm, extract representations enriched with both global (e.g., coarse structural textures and global pathological severity) and local (e.g., cellular morphology, lesion localization, and fine-grained textures) cues from pathology images, including H&E. These implicit yet powerful representations, though difficult to observe directly or define explicitly, exhibit remarkable generalization in downstream tasks such as classification, segmentation, and detection. Research also demonstrates their potential in generative tasks such as histopathology image generation Yellapragada et al. (2025a).

Inspired by these insights, we hypothesize that the implicit representations extracted by pathology foundation models from H&E images can provide rich pathological and structural priors for IHC synthesis in virtual staining (validated in Table 4). However, their utilization remains underexplored. Most GAN- and diffusion-based approaches rely on explicit representations, such as classification labels Jia et al. (2025); Li et al. (2024b), optical density maps Chen et al. (2024a), frequency information Zhang et al. (2024), and cell segmentation masks Kataria et al. (2024). HistDiST Großkopf et al. (2025) leverages Phikon Filiot et al. (2023) to capture morphological structures, whereas PixCell Yellapragada et al. (2025a) employs UNI2-h Chen et al. (2024b) for embedding to synthesize IHC images. However, both use only the global class tokens from the foundation models, ignoring the rich, fine-grained information in the abundant patch tokens.

To better exploit the rich representations from pathology foundation models, we propose **KIIF**, a **k**nowledge-**i**ntensive **i**mplicit representations **f**usion module that leverages long-sequence visual tokens, including all global and patch tokens, to enhance H&E-to-IHC translation. The global tokens provide high-level conditions, such as cancer severity, whereas the patch tokens inject local, fine-grained information, including morphology, localization, and textures. Observing that various pathology foundation models are trained on different data with distinct strategies, we further hypothesize that they may offer complementary guidance for virtual staining. In this work, we experiment with combining the outputs of up to five foundation models as a pilot exploration. Notably, incorporating such a significant number of tokens for attention is enabled by the DC-AE and linear-attention-based flow matching network, which effectively reduces inference latency and memory demand.

In summary, our contributions are as follows:

- We introduce StainFlash, a DC-AE and linear-attention-based flow matching model that enables high-resolution, high-quality H&E-to-IHC virtual staining with significantly reduced inference latency and memory demand.
- We propose a lightweight pathology encoder that extracts layout embeddings from H&E images, dependent on the timestep in the flow matching process.
- We design the KIIF module that incorporates both global and fine-grained implicit representations from dual pathology foundation models, ensuring high-quality and effective translation of pathological information.
- Last but not least, we conduct extensive experiments on four datasets to validate StainFlash against state-of-the-art methods. We also assess a downstream task and perform ablation studies on StainFlash.

2 Related Work

2.0.1 GAN-based Virtual Staining

Early approaches primarily relied on GANs for virtual staining. Li et al. (2024a); Wang et al. (2024a) targeted H&E-to-IHC conversion. Other studies Asaf et al. (2024); Rivenson et al. (2019) demonstrated the synthesis of H&E images from formalin-fixed, paraffin-embedded (FFPE) tissue. Most GAN-based methods incorporated explicit pathology representations—such as classification labels Jia et al. (2025), optical density maps Chen et al. (2024a), frequency information Zhang et al. (2024), or diaminobenzidine (DAB) channel supervision Wei et al. (2024)—for more accurate transfer of pathological semantics. Others utilized explicit structural representations to guide the translation, such as H-channel similarity Wei et al. (2024), structural representations derived from external tools Wang et al. (2024a), and auxiliary structural cues Dubey et al. (2023); Pati et al. (2024); Wang et al. (2024b) (e.g., Canny edge maps Dubey et al. (2023)), to enforce consistency between source and target images concerning cellular morphology, tissue architecture, stromal texture, and spatial organization. Several studies further disentangled pathological and structural representations Wang et al. (2025b); Yuan et al. (2025), optimizing them separately for better controllability and fidelity. Nevertheless, implicit representations embedded by large-scale pathology foundation models remain largely underexplored for virtual staining. Moreover, GAN-based methods suffer from mode collapse and unstable training dynamics.

Diffusion-based Virtual Staining Various studies explored diffusion models Esser et al. (2024) for virtual staining in the pixel space Chen et al. (2025c). They also introduced explicit representations—e.g., frequency-domain information He et al. (2024) and cell segmentation masks Kataria et al. (2024)—or attempted to leverage implicit representations from pathology vision-language models (VLMs; Chen et al., 2025c) to enhance virtual staining. However, due to substantial computational and memory overhead, these methods are typically restricted to low-resolution synthesis at 256×256 pixels. Recent studies Dubey et al. (2024) built upon latent-space generative frameworks, enabling high-resolution (e.g., 512×512 or even 1024×1024 pixels) virtual staining with significantly reduced computational costs. Großkopf et al. (2025) investigated integrating implicit representations from pathology foundation models Chen et al. (2024b) via cross-attention. However, due to the quadratic complexity of attention mechanisms with respect to sequence length, these methods encounter substantial inference latency at high resolution and use only a small subset of the implicit representation tokens.

Diffusion Model Acceleration To accelerate diffusion model inference, various methods explored step distillation (e.g., consistency model training Luo et al. (2023a), adversarial learning Sauer et al. (2024), and variational score distillation), albeit at the cost of additional training overhead and suboptimal quality. Others leveraged quantization, pruning, and feature map reuse Liu et al. (2025b) to reduce per-step computational cost, yet remain constrained by quadratic attention complexity. Going further, Fei et al. (2024); Peruzzo et al. (2025) compressed token representations or employed local attention Liu et al. (2024a), yet still suffered from sub-quadratic attention overhead. Liu et al. (2024b); Phung et al. (2024); Zhu et al. (2025) explored linear-complexity attention in diffusion/flow matching models via Mamba or linear attention. These advances, however, have yet to be demonstrated effective for pathology image generation and virtual staining. We address this gap with a flow matching network

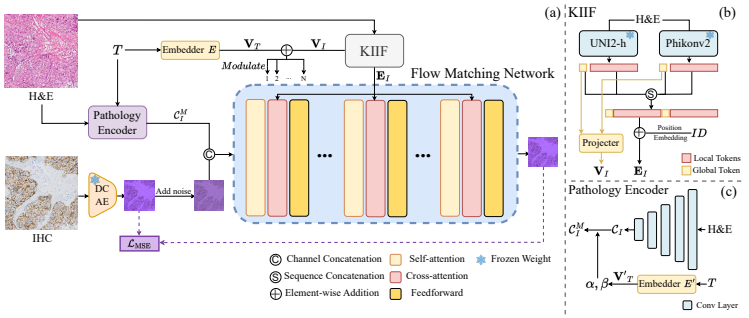

Figure 1: (a) **Framework overview**. StainFlash employs a DC-AE and linear-attention-based flow-matching network for efficient H&E-to-IHC virtual staining. It also integrates a KIIF module, which fuses the intensive implicit representations extracted by dual pathology foundation models, and a pathology encoder to extract timestep-modulated layout embeddings from H&E images. (b) The **KIIF** module. (c) The **pathology encoder**. Best viewed zoomed-in.

using a deep-compression autoencoder and linear attention, achieving high-resolution virtual staining of pathology images efficiently.

3 METHOD

3.0.1 PRELIMINARIES:

Flow matching Liu et al. (2022b) has emerged as an appealing alternative to diffusion models, due to its high sampling speed and competitive sample quality. It belongs to the family of flow-based generative models, which estimate a transformation (a flow) between a prior distribution (e.g., Gaussian) and a target distribution. Unlike normalizing flows, which learn direct noise-to-data transformations under fixed architectural constraints (e.g., affine coupling), flow matching regresses the time-dependent vector field that generates this flow by solving the corresponding ordinary differential equations with flexible architectures.

For training, given an image latent z_1, a random noise $z_0 \sim \mathcal{N}(0, I)$, and a timestep $t \in [0, 1]$ sampled from a logit-normal distribution, an intermediate latent z_t is obtained as the training input. Following rectified flows Esser et al. (2024), z_t is defined as a linear interpolation between z_0 and z_1, and the ground truth velocity v_t is its derivative with respect to t:

$$z_t = tz_1 + (1-t)z_0, \quad v_t = dz_t/dt = z_1 - z_0. \tag{1}$$

The model is trained to predict the velocity. Consequently, its loss function is defined as the mean squared error (MSE) calculated between the model output and v_t:

$$\mathcal{L} = \mathbb{E}_{z_0, z_1, c, t} \left\| u(z_t, c, t; \theta) - v_t \right\|^2, \tag{2}$$

where c is the condition, θ denotes model parameters, and $u(z_t, c, t; \theta)$ is the predicted velocity. In our context, c encompasses the implicit representations from pathology foundation models and the layout embedding encoded by the pathology encoder, both derived from H&E images.

3.0.2 METHOD OVERVIEW

Our target is to translate images from the H&E domain $\mathcal{X} \subset \mathbb{R}^{H \times W \times 3}$ to the IHC domain $\mathcal{Y} \subset \mathbb{R}^{H \times W \times 3}$, 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 IHC image $\hat{y}$ should exhibit high visual quality (e.g., clear tissue; and color distributions that align well with the IHC domain) and high pathological fidelity (textural alignment in nuclear distribution and sizes, and pathological consistency in cancer severity with y).

Figure 1(a) overviews StainFlash. A lightweight pathology encoder encodes the H&E image into a layout embedding $\mathbf{C}_I^M$, which is concatenated channel-wise with the noise latent and injected into the flow matching network. Meanwhile, the **KIIF** module extracts knowledge-intensive implicit

representations, including both global and patch tokens that contain pathological and textual priors, from the H&E image using two pathology foundation models. The global tokens provide high-level conditions. They are projected to vectors $\mathbf{V}_I$, which, together with the time step embedding $\mathbf{V}_T$, modulate the flow matching network with learned timestep-dependent weights. The patch tokens $\mathbf{E}_I$ are fused via linear cross-attention into the flow matching network, providing local priors for localized conditioning.

Pathology Encoder with Timestep-dependent Modulation As shown in Figure 1(c), the pathology encoder consists of a lightweight convolutional network and a timestep-dependent modulation branch. The convolutional network (comprising five convolutional layers with 4×4 kernels, 2×2 strides, and channel sizes of 8, 32, 64, 256, and 1024) encodes the H&E image into layout embeddings $\mathbf{C}_I$ of the same shape as the noise latent.

Studies show that diffusion and flow matching Esser et al. (2024) exhibit a temporal-hierarchical division of labor: early timesteps reconstruct coarse structures while later ones refine high-frequency details. Therefore, directly concatenating $\mathbf{C}_I$ with the noise latent—i.e., applying identical conditioning across all timesteps—risks suboptimally fusing the H&E embedding into the flow matching network. We employ a time embedder E' to project the global timestep $T = t \times 1000$ into a sinusoidal embedding $\mathbf{V}'_T$. $\mathbf{V}'_T$ is then projected into channel-wise scale α and bias β by a time-dependent modulation projector $\mathcal{M}$ (composed of a linear layer and a SiLU activation). α and β modulate $\mathbf{C}_I$ to produce the modulated layout embedding $\mathbf{C}_I^M$:

$$\alpha, \beta = \mathcal{M}(\mathbf{V}'_T), \quad \mathbf{C}_I^M = (1 + \alpha) \cdot \mathbf{C}_I + \beta. \tag{3}$$

$\mathbf{C}_I^M$ is concatenated channel-wise with the noise latent and fed to the flow matching network. The norms of α and β across timesteps reveal that, at larger timesteps (early stage), $\mathbf{C}_I$ injects stronger priors from abundant structural cues, whereas at smaller timesteps (late stage) the injection becomes more conservative, facilitating fine-grained detail refinement (supplementary Figure S2). The ablation studies (cf. Table 4) suggest that the timestep-dependent modulation of $\mathbf{C}_I$ facilitates better fusion of H&E features and leads to notable improvements in generation quality.

3.0.3 KNOWLEDGE-INTENSIVE IMPLICIT REPRESENTATIONS FUSION (KIIF)

Recent pathology foundation models Chen et al. (2024b), built upon the Vision Transformer (ViT; Dosovitskiy, 2020), are capable of extracting representations enriched with both global and local cues from pathology images, including H&E images. In ViT, an image is divided into patches, each encoded as a patch token preserving spatial position and local semantics. Global information is captured by a learnable global token that interacts with all patch tokens via attention and captures holistic semantics of the entire image. Both the patch and global tokens constitute implicit representations—though difficult to observe directly or define explicitly, e.g., in classification labels, optical density maps, or frequency information—yet are effective for various downstream and generative tasks. However, their potential in virtual staining remains largely underexplored. While Großkopf et al. (2025) have made preliminary attempts, they only exploit the global tokens, omitting the abundant spatial and local semantics of patch tokens.

To fully harness the knowledge-intensive implicit representations from foundation models, we propose the KIIF module (Figure 1(b)). KIIF leverages two representative and powerful foundation models, UNI2-h Chen et al. (2024b) and Phikonv2 Filiot et al. (2024), for complementary prior extraction. To deeply integrate global prior information, similar to Kong et al. (2024), we combine the timestep embedding $\mathbf{V}_T$ (encoded by a time embedder E) and the global tokens from both models as inputs to the modulation mechanism. Specifically, the two global tokens are projected through independent linear layers and summed element-wise to form the mixed image embedding $\mathbf{V}_I$, which is then added element-wise to $\mathbf{V}_T$. Similar to Eqn. (3), the aggregated representation is fed into a modulation projector, producing the modulation factors α and β for each block of the flow matching network. The norm ratio between $\mathbf{V}_I$ and $\mathbf{V}_T$ demonstrates that $\mathbf{V}_I$ exerts a substantial modulation effect on the overall flow matching inference process (supplementary Figure S3).

However, the global representation predominantly captures high-level, holistic information of the H&E image, such as general textures and global pathological severity. To complement it, StainFlash incorporates fine-grained, localized representations into the flow matching network. All global and patch tokens from UNI2-h and Phikonv2 are concatenated along the sequence dimension. Meanwhile,

Table 1: Evaluation of various methods on the (a) RegH2I, (b) RegH2I-Ext, (c) BCI, (d) MIST-HER2, and (e) MIST-ER datasets.

Method	(a) RegH2I Peng et al. (2024)			(b) RegH2I-Ext Peng et al. (2024)			(c) BCI Liu et al. (2022a)			(d) MIST-HER2 Li et al. (2023)			(e) MIST-ER Li et al. (2023)		
	DISTS↓	FID↓	KID↓	DISTS↓	FID↓	KID↓	DISTS↓	FID↓	KID↓	DISTS↓	FID↓	KID↓	DISTS↓	FID↓	KID↓
CycleGAN	0.212	40.91	0.0062	0.282	149.6	0.1168	0.262	61.09	0.0304	0.254	67.78	0.0343	0.265	46.66	0.0142
USIGAN	0.262	55.34	0.0249	0.248	128.1	0.0670	0.310	96.85	0.0752	0.230	37.21	0.0049	0.233	33.59	0.0037
AI-FFPE	0.221	37.15	0.0047	0.287	92.89	0.0365	0.301	51.31	0.0226	0.258	57.41	0.0217	0.267	72.54	0.0413
Pix2pix	0.192	47.77	0.0237	0.292	147.0	0.1065	0.255	91.06	0.0580	0.258	75.97	0.0427	0.256	70.53	0.0382
Pix2pixHD	0.191	41.77	0.0123	0.279	104.9	0.0597	0.280	88.81	0.0583	0.249	61.45	0.0236	0.251	58.45	0.0213
Peng et al. 2024	0.174	33.92	0.0058	0.265	104.7	0.0796	0.271	63.18	0.0258	0.238	50.95	0.0148	0.238	54.68	0.0204
pix2pix-Turbo	0.451	225.2	0.2909	0.503	289.2	0.3261	0.489	303.1	0.4019	0.465	152.5	0.1018	0.418	79.73	0.0371
ControlNet	0.383	209.7	0.2560	0.374	235.9	0.2741	0.394	236.9	0.2381	0.343	170.2	0.1244	0.347	245.6	0.2156
StainFuser	0.257	123.7	0.1057	0.249	182.7	0.1438	0.275	56.67	0.0268	0.241	58.36	0.0228	0.248	62.79	0.0322
HistDiST	0.257	70.99	0.0389	0.273	110.2	0.0526	0.254	56.12	0.0230	0.284	77.53	0.0408	0.301	95.75	0.0654
PixCell	0.220	47.77	0.0138	0.268	102.1	0.0387	**0.201**	42.24	0.0179	0.214	50.23	0.0185	0.207	43.88	0.0166
StainFlash (ours)	**0.173**	**31.66**	**0.0033**	**0.241**	**79.77**	**0.0318**	0.222	**32.00**	**0.0077**	**0.208**	**32.46**	**0.0020**	**0.197**	**27.81**	**0.0017**

each model ID (UNI2-h = 1 and Phikonv2 = 2) is encoded into a model-specific positional, sinusoidal embedding. Then, the embedding is element-wise added to its corresponding patch and global tokens, producing the final sequence representation $\mathbf{E}_I$. $\mathbf{E}_I$ is injected into all blocks of the flow matching network via cross-attention, providing comprehensive H&E priors from dual foundation models.

Linear Attention and DC-AE Enabled KIIF Despite providing comprehensive H&E priors, concatenating all global and patch tokens of two foundation models produces a long token sequence, for which traditional $\mathcal{O}(N^2)$ softmax attention incurs prohibitive computational cost and inference latency. Unlike SANA Xie et al. (2025) and EDiT Becker et al. (2025), which use softmax attention for cross-modal fusion, we replace all softmax attention in the flow matching network with linear attention, enabling efficient long-sequence fusion. Ablation studies show comparable performance to $\mathcal{O}(N^2)$ softmax attention with considerably improved efficiency (Tables 2 and 4).

Consider a standard softmax attention with h heads, sequence length N, and hidden dimension $D = h \cdot d$. Let $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{N \times D}$ denote the query, key, and value. The output $\mathbf{O} \in \mathbb{R}^{N \times D}$ can be expressed as:

$$\mathbf{O}_i = \sum\nolimits_{j=1}^{N} \Big[\mathrm{Sim}(\mathbf{Q}_i, \mathbf{K}_j) \Big/ \sum\nolimits_{j=1}^{N} \mathrm{Sim}(\mathbf{Q}_i, \mathbf{K}_j) \Big] \mathbf{V}_j, \tag{4}$$

where $\mathrm{Sim}(\mathbf{Q}, \mathbf{K}) = \exp(\mathbf{Q}\mathbf{K}^\top / \sqrt{d})$ is a similarity measurement function. Linear attention simplified the function above with a nonlinear kernel $\boldsymbol{\phi}(\cdot)$, *i.e.*, $\mathrm{Sim}(\mathbf{Q}, \mathbf{K}) = \boldsymbol{\phi}(\mathbf{Q})\boldsymbol{\phi}(\mathbf{K})^\top$. Thus, Eqn. (4) can be rewritten as Eqn. (5) and simplified to Eqn. (6) using the associative property of multiplication Wang et al. (2025a):

$$\mathbf{O}_i = \sum\nolimits_{j=1}^{N} \frac{\boldsymbol{\phi}(\mathbf{Q}_i)\, \boldsymbol{\phi}(\mathbf{K}_j)^\top}{\sum_{j=1}^{N} \boldsymbol{\phi}(\mathbf{Q}_i)\, \boldsymbol{\phi}(\mathbf{K}_j)^\top} \mathbf{V}_j, \tag{5} \qquad \mathbf{O}_i = \frac{\boldsymbol{\phi}(\mathbf{Q}_i) \Big(\sum_{j=1}^{N} \boldsymbol{\phi}(\mathbf{K}_j)^\top \mathbf{V}_j \Big)}{\boldsymbol{\phi}(\mathbf{Q}_i) \Big(\sum_{j=1}^{N} \boldsymbol{\phi}(\mathbf{K}_j)^\top \Big)}. \tag{6}$$

The computational complexities of the softmax and linear attention are $\mathcal{O}(N^2 D)$ and $\mathcal{O}(N D^2 / h)$, respectively.

Meanwhile, the image-side latent representation also contributes significantly to the computational cost of the flow matching network. To mitigate this, StainFlash adopts the DC-AE Chen et al. (2025b) from SANA Xie et al. (2025), which uses a $32\times$ spatial compression ratio (vs. $8\times$ in conventional autoencoders Esser et al. (2024)). This shortens the image-side token sequences by a factor of 16, significantly reducing latency and memory consumption (Table 2). DC-AE maintains high reconstruction quality despite this aggressive compression via non-uniform downsampling with residual blocks and an expanded latent channel count.

4 Experiments

4.0.1 Datasets and Evaluation Metrics

We conduct extensive experiments on four H&E-to-IHC virtual staining datasets: RegH2I Peng et al. (2024), BCI Liu et al. (2022a), MIST-HER2 Li et al. (2023), and MIST-ER Li et al. (2023); details in Section B. Following existing practice Yang et al. (2025), we report the deep image structure and texture similarity (DISTS) Ding et al. (2020), Fréchet inception distance (FID) Heusel et al. (2017), and kernel inception distance (KID) Bińkowski et al. (2018) for a comprehensive evaluation. These metrics are known to align with human perceptual quality (lower is better).

Implementation Experiments are conducted on Ubuntu 20.04, Python 3.10.16, and PyTorch 2.4.0, on an 80 GB GPU. We use SANA-0.6B Xie et al. (2025) as the backbone for the flow matching network, initialized from pretrained SANA_600M_1024px_MultiLing. We use the CAME-AdamW optimizer Luo et al. (2023b) with a constant learning rate of 5×10^{-4}, betas $(0.9, 0.999, 0.9999)$, and eps $(1.0\text{e}{-30}, 1.0\text{e}{-16})$. The model is trained for 30,000 steps on 1024×1024-pixel crops with a batch size of 8, with a 2,000-step learning rate warm-up. Our code and trained models will be released.

Comparison with State-of-the-Art (SOTA) We compare StainFlash with: (i) classical GANs—CycleGAN Zhu et al. (2017), USIGAN Peng et al. (2026), pix2pix Isola et al. (2017), and pix2pixHD Wang et al. (2018); (ii) SOTA diffusion models—ControlNet Zhang et al. (2023) and pix2pix-Turbo Parmar et al. (2024); and (iii) specialized virtual staining approaches—GAN-based: Peng et al. (2024) and AI-FFPE Ozyoruk et al. (2021), diffusion-based: StainFuser Jewsbury et al. (2024) and HistDiST Großkopf et al. (2025), and flow matching-based: PixCell Yellapragada et al. (2025a). For multi-step diffusion/flow matching models (ControlNet, StainFuser, HistDiST, and PixCell), we follow their default sampling steps (50, 20, 200, and 20, respectively). StainFlash uses 20 sampling steps throughout.

Quantitative results. Table 1 reports the evaluation results. GANs generally outperform diffusion models on this challenging task, though the best diffusion model, StainFuser, is broadly competitive with many GANs. Overall, Peng et al. (2024) and PixCell Yellapragada et al. (2025a) represent the previous SOTA, achieving the best numbers for most metrics across datasets. StainFlash achieves the best performance on nearly all metrics across all four datasets (second-best on DISTS for BCI), establishing a new SOTA for H&E-to-IHC virtual staining. These results confirm both the superior quality of StainFlash's virtually stained IHC images and its generalizability across metrics and datasets. We further report CONCH-based Lu et al. (2024) cosine feature similarity and FID in supplementary Tables S2 and S4. CONCH is a histopathology-specific vision-language model; its features may provide a more pathology-oriented assessment than those from natural-image pretrained extractors. StainFlash consistently ranks best on both metrics across all datasets, demonstrating excellent performance and generalizability.

Qualitative analysis. We further present qualitative results for all four datasets to support in-depth analysis. Figure 2 shows results on RegH2I Peng et al. (2024) (BCI Liu et al. (2022a), MIST-HER2, and MIST-ER Li et al. (2023) results are in supplementary Figures S6–S8). StainFlash outperforms prior methods in both perceptual realism and visual plausibility, producing clear and sharp nuclei and colors and textures that most closely match real IHC images. In contrast, HistDiST Großkopf et al. (2025) and StainFuser Jewsbury et al. (2024) yield over-smoothed membrane textures in lesion regions (last two rows) and nuclei that appear either overly dark or too faint (first two rows). PixCell Yellapragada et al. (2025a) produces artifactual colors in nuclei and membranes with reduced visual plausibility (particularly in the last three rows). AI-FFPE Ozyoruk et al. (2021) and CycleGAN Zhu et al. (2017) also generate blurred nuclei in the second row.

StainFlash also exhibits higher textural and pathological fidelity: its nuclear distributions and sizes best match the ground truth. Other methods either underestimate (pix2pix-Turbo Parmar et al. (2024), AI-FFPE Ozyoruk et al. (2021), Peng et al. (2024), Pix2pix Isola et al. (2017); last two rows) or overestimate (ControlNet Zhang et al. (2023), CycleGAN Zhu et al. (2017); first two rows) cancer presence/severity, compounded by imprecise lesion localization. These align with StainFlash's superior performance in Table 1. The supplementary provides additional examples (including a visual failure analysis in Figure S4). Two board-certified pathologists blindly ranked virtual staining results of several whole-slide images from 1 (best) to 4 (worst) by perceptual quality and pathological fidelity. StainFlash ranks first among four methods, demonstrating superior scalability and staining quality. These encouraging perceptual, visual, and pathological qualities imply the effectiveness of incorporating intensive implicit pathology representations from dual foundation models into the flow matching network.

Complexity Analysis We benchmark end-to-end inference latency and memory consumption for generating an image of 1024×1024 pixels. We compare StainFlash with the two best diffusion/flow matching methods in Table 1, StainFuser Jewsbury et al. (2024) and PixCell Yellapragada et al. (2025a). Notably, PixCell uses a flow matching network of comparable size to StainFlash, yet operates with an $8\times$ downsampled latent space and softmax attention. It further applies a $2\times$

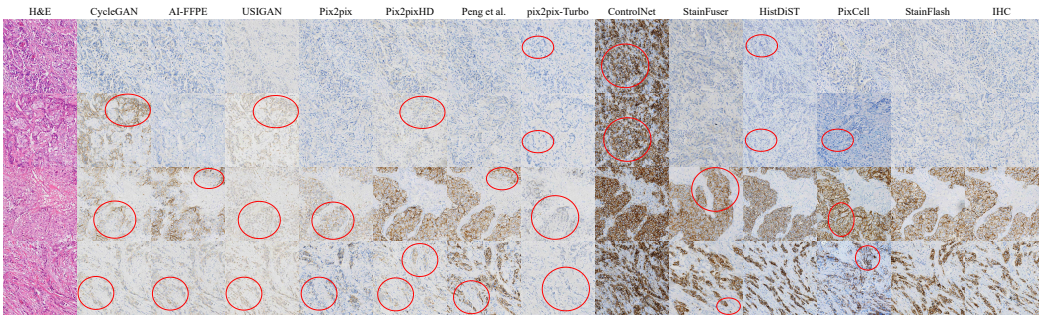

Figure 2: Example virtual staining results on RegH2I. In IHC images, blue dots denote cell nuclei, and brown regions in the cell membranes correspond to cancerous lesions, with darker brown coloration signifying more advanced lesion severity and further disease progression. Red ellipses highlight regions where compared methods yield erroneous nuclei, colors, textures, or lesion severity while ours are correct. Best viewed zoomed-in.

Table 2: Evaluation of computational complexity on an 80-GB GPU in inference latency and memory consumption.

Method	Type	Parameters	Downsample	Attention	Latency	Memory
Peng et al. 2024	GAN	45.59M	-	-	0.013s	2.574G
StainFuser	Diffusion	1,230M	8×	Softmax	3.689s	10.322G
PixCell	Flow matching	608.0M	16×	Softmax	4.457s	9.059G
StainFuser*	Diffusion	1,230M	**32**×	Softmax	1.741s	10.05G
PixCell*	Flow matching	608.0M	**32**×	Softmax	4.195s	8.962G
StainFlash-8-SM	Flow matching	607.9M	8×	Softmax	12.590s	10.071G
StainFlash-32-SM	Flow matching	607.9M	32×	Softmax	1.325s	8.782G
StainFlash	Flow matching	607.8M	32×	Linear	**1.149**s	**8.781**G

* Benchmark against StainFlash under the 32× downsampling, instead of the original ratios of compared methods.

Table 3: Evaluation of a downstream classification task.

Method	ACC↑	F1↑
H&E	0.4207±0.0152	0.3170±0.0141
Real IHC	0.7270±0.0034	**0.7320**±0.0032
CycleGAN	0.5333±0.0125	0.4781±0.0116
AI-FFPE	0.6972±0.0083	0.6304±0.0075
USIGAN	0.5867±0.0107	0.5729±0.0101
Peng et al. 2024	0.7179±0.0086	0.6698±0.0083
Pix2pix	0.5895±0.0103	0.5226±0.0097
Pix2pixHD	0.6456±0.0106	0.5487±0.0100
ControlNet	0.4021±0.0151	0.3267±0.0145
HistDiST	0.7119±0.0082	0.6872±0.0078
StainFuser	0.5888±0.0100	0.5356±0.0097
pix2pix-Turbo	0.3554±0.0152	0.2106±0.0145
PixCell	0.6502±0.0084	0.6168±0.0076
StainFlash (ours)	**0.7305**±0.0039	0.7137±0.0036

Table 4: Ablation study on the RegH2I dataset.

Method	DISTS↓	FID↓	KID↓
w/o Implicit	0.188	36.39	0.0098
w/o UNI	0.176	33.83	0.0097
w/o Phi	0.175	33.75	0.0052
w/o UNI-G	0.173	33.26	0.0063
w/o UNI-P	0.181	34.44	0.0065
w/o Phi-G	0.175	32.19	0.0075
w/o Phi-P	0.174	33.05	0.0057
w/o Linear	**0.171**	31.83	0.0036
w/o $\mathbf{V}'_T$	0.176	33.69	0.0056
w/o $\mathbf{V}_I$	0.177	33.53	0.0056
w/o *ID*	0.176	33.79	0.0082
StainFlash (ours)	0.173	**31.66**	**0.0033**

downsampling via convolution to latents before attention—*effectively using 16× downsampling*. The best GAN-based method Peng et al. (2024) is included, too.

Table 2 shows that StainFlash uses only 31% of StainFuser's latency and 85% of its memory. PixCell requires 287% more time and 3% more memory than StainFlash, despite comparable parameters. These substantial gains demonstrate StainFlash's efficiency advantages over SOTA diffusion/flow matching methods, beyond its superior virtual staining quality. We further benchmark StainFuser and PixCell under 32× downsampling (matching StainFlash), instead of their original ratios. StainFlash consistently achieves lower latency and reduced memory at the same ratio, reflecting its architectural advantages.

Among variants of StainFlash, 32× downsampling (StainFlash-32-SM) cuts latency by 89% and memory by 13% over 8× (StainFlash-8-SM). When scaling up to a batch size of 8, the memory saving increases to 27%. While 8× downsampling may yield better staining quality than 32× (cf. supplementary Table S8), StainFlash balances superior performance and computational efficiency over SOTA diffusion/flow matching methods—a favorable quality-efficiency trade-off. Replacing softmax with linear attention further reduces latency by 13%, underscoring major efficiency benefits. The GAN-based method Peng et al. (2024) is orders of magnitude faster and more memory-efficient than diffusion/flow matching-based ones, due to the different nature of its methodology. Nevertheless, StainFlash consistently outperforms it across diverse datasets and metrics.

External Validation To assess generalizability, we further validate StainFlash on the external test dataset (RegH2I-Ext) from Peng et al. (2024). This dataset comprises 285 paired H&E-IHC images stained with SP3 D'Alfonso et al. (2013) or CB11 Purdie et al. (2010) antibodies. We directly apply the models trained on RegH2I in the main experiments without fine-tuning. As shown in Table 1(b) , StainFlash consistently outperforms other methods across all three metrics on the independent patient cohorts. These results align well with the main experiments in Table 1, demonstrating StainFlash's strong generalizability for H&E-to-IHC virtual staining.

Evaluation on Downstream Task To assess the potential clinical value of virtually stained IHC images, we conduct a membrane staining intensity classification study on RegH2I and RegH2I-Ext. The experiment adheres to the HER2 scoring criteria Ahn et al. (2020), involving a four-level classification. For each method, 600 virtual IHC patches generated from the H&E test images in RegH2I are used to train a ResNet-50 classifier, which is then assessed on real IHC images from RegH2I-Ext, via accuracy and F1 score. Each setting is repeated over 10 runs, and the mean and standard deviation are reported. For reference, we also train classifiers on H&E images and real IHC images from the RegH2I test set. As shown in Table 3, the model trained on IHC images synthesized by our approach outperforms those based on other methods, and approaches the upper-bound classifier trained on real IHC images. The per-class F1 scores are: ours 0.74/0.62/0.58/0.91 vs. real IHC 0.73/0.69/0.63/0.87 for HER2 0/1+/2+/3+, with larger gaps in moderate-positive regions (1+/2+). The H&E-trained classifier shows limited performance, as expected, as it cannot capture membrane staining intensity.

Ablation Studies We ablate key components of StainFlash on RegH2I to assess their contributions (Table 4). We hypothesize that: (H1) the intensive implicit representations from pathology foundation models benefit H&E-to-IHC virtual staining, and (H2) different foundation models provide complementary guidance. To test the hypotheses, we evaluate: 1) removing all implicit representations by substituting them with zero embeddings (denoted by "w/o Implicit"); 2) removing either UNI2-h ("w/o UNI") or Phikonv2 ("w/o Phi"); 3) removing the global token of either ("w/o UNI-G", "w/o Phi-G"); and 4) removing the patch tokens of either ("w/o UNI-P", "w/o Phi-P"). Results show that removing implicit representations entirely yields substantial deterioration across DISTS, FID, and KID, validating H1. Removing either UNI2-h or Phikonv2 yields notable degradation, confirming H2. Removing either global or patch tokens from either model causes moderate degradation, indicating that both token types constitute intensive implicit representations and are essential for optimal performance.

Supplementary Tables S6 and S7 ablate combinations of up to five pathology foundation models; the UNI2-h-Phikonv2 pair offers a favorable performance-efficiency balance. Most dual-model combinations are competitive with leading methods in Table 1, indicating that StainFlash flexibly incorporates alternative foundation models for consistently strong performance. StainFlash with five models (1.458 s) remains faster than methods compared in Table 2.

The long sequence of knowledge-intensive implicit representation is fused into the flow matching network through cross-attention. To accelerate long-sequence cross-attention, we replace all conventional softmax attention with linear attention, reducing complexity from $\mathcal{O}(N^2)$ to $\mathcal{O}(N)$. For ablation, we revert to softmax attention ("w/o Linear"). Softmax and linear attention perform comparably, yet the latter notably reduces inference latency (Table 2).

We further ablate: 1) timestep-dependent modulation in the pathology encoder ("w/o $\mathbf{V'}_T$"); 2) global-token modulation of the flow matching network ("w/o $\mathbf{V}_I$"); and 3) positional embeddings for foundation models ("w/o *ID*"). All three ablations cause considerable degradation, confirming their importance for effective multi-source fusion.

StainFlash adopts pretrained SANA Xie et al. (2025) as its backbone, following common practice in diffusion and flow matching frameworks that leverage models pretrained on large-scale natural image datasets for improved initialization and performance. We evaluate two alternatives: (i) training from scratch, and (ii) freezing pretrained SANA weights. As shown in supplementary Table S9, training from scratch yields substantially worse performance, likely due to the limited virtual-staining samples for training a flow matching network from scratch. Freezing the pretrained weights also leads to inferior results. These findings confirm that, while SANA pretraining provides a strong initialization, the performance gain stems primarily from our framework design rather than pretraining alone.

5 Conclusion

This work introduces StainFlash, an efficient framework for high-resolution H&E-to-IHC virtual staining. Integrating DC-AE and linear-attention-based flow matching overcomes/relieves the latency and memory constraints of prior methods. A novel KIIF module leverages rich, implicit representations from multiple pathology foundation models. Extensive experiments demonstrate superior performance to existing approaches. Limitations and future work are discussed in Section E.

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 StainFlash, an application-oriented framework for efficient and high-fidelity H&E-to-IHC 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. StainFlash 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, StainFlash 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.

REPRODUCIBILITY STATEMENT

We are committed to ensuring the reproducibility of StainFlash. 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 4.0.1 (Experiments) of the main paper. The network architecture, the pathology encoder with timestep-dependent modulation, the KIIF module, and the flow matching objective are described in Section 3.0.1 (Method) and the subsequent method subsections. Additional experimental settings, including ablation studies on pathology foundation models, downsampling ratios, pretrained weights, and fine-tuning of the autoencoder, are provided in Appendix Tables S6, S8, S9, and S10.

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. *J. Pathol. Transl. Med.*, 54(1):34–44, 2020.

Muhammad Zeeshan Asaf et al. 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 *ICML*. JMLR, 2023.

Philipp Becker et al. Edit: Efficient diffusion transformers with linear compressed attention. In *ICCV*, pp. 19608–19616, 2025.

Mahesh Bhosale et al. Pathdiff: Histopathology image synthesis with unpaired text and mask conditions. *arXiv preprint arXiv:2506.23440*, 2025.

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. 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 *MICCAI*, pp. 384–394. Springer, 2024a.

Junsong Chen et al. Sana-sprint: One-step diffusion with continuous-time consistency distillation. *ICCV*, 2025a.

Junyu Chen et al. Deep compression autoencoder for efficient high-resolution diffusion models. 2025b.

Richard J Chen et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 2024b.

Zizhi Chen et al. VLM-based prompts as the optimal assistant for unpaired histopathology virtual staining. In *ACM MM*, pp. 3527–3536, 2025c.

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. *J. Clin. Pathol.*, 66(5):409–414, 2013.

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

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 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.

Patrick Esser et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024.

Zhengcong Fei, Mingyuan Fan, Changqian Yu, Debang Li, and Junshi Huang. Diffusion-rwkv: Scaling rwkv-like architectures for diffusion models. *arXiv preprint arXiv:2404.04478*, 2024.

Alexandre Filiot, Paul Jacob, Alice Mac Kain, and Charlie Saillard. Phikon-v2, a large and public feature extractor for biomarker prediction, 2024.

Alexandre Filiot et al. Scaling self-supervised learning for histopathology with masked image modeling. *MedRxiv*, pp. 2023–07, 2023.

Ian J Goodfellow et al. Generative adversarial nets. *NIPS*, 27, 2014.

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

Yufang He et al. PST-Diff: achieving high-consistency stain transfer by diffusion models with pathological and structural constraints. *IEEE TMI*, 2024.

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. *NIPS*, 30: 6629–6640, 2017.

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 *WACV*, pp. 4382–4391. IEEE, 2025.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 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.

kaiko.ai et al. Towards large-scale training of pathology foundation models, 2024.

Mingu Kang, Heon Song, Seonwook Park, Donggeun Yoo, and Sérgio Pereira. Benchmarking self-supervised learning on diverse pathology datasets. In *CVPR*, pp. 3344–3354, June 2023.

Mikhail Karasikov et al. Training state-of-the-art pathology foundation models with orders of magnitude less data. In *MICCAI*, pp. 573–583. Springer, 2025.

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.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *ICML*, pp. 5156–5165. PMLR, 2020.

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.

Weijie Kong et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 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 *MICCAI*, pp. 632–641. Springer, 2023.

Jiahan Li et al. Virtual immunohistochemistry staining for histological images assisted by weakly-supervised learning. In *CVPR*, pp. 11259–11268, 2024a.

Yuzhu Li 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. In *International Workshop on Computational Mathematics Modeling in Cancer Analysis*, pp. 30–39. Springer, 2025a.

Feng Liu et al. Timestep embedding tells: It's time to cache for video diffusion model. In *CVPR*, pp. 7353–7363, 2025b.

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

Songhua Liu, Zhenxiong Tan, and Xinchao Wang. Clear: Conv-like linearization revs pre-trained diffusion transformers up. *arXiv preprint arXiv:2412.16112*, 2024a.

Songhua Liu, Weihao Yu, Zhenxiong Tan, and Xinchao Wang. Linfusion: 1 gpu, 1 minute, 16k image. *arXiv preprint arXiv:2409.02097*, 2024b.

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

Ming Y Lu et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30:863–874, 2024.

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*, 2023a.

Yang Luo, Xiaozhe Ren, Zangwei Zheng, Zhuo Jiang, Xin Jiang, and Yang You. CAME: Confidence-guided adaptive memory efficient optimization. In *ACL*, pp. 4442–4453, 2023b.

Dmitry Nechaev, Alexey Pchelnikov, and Ekaterina Ivanova. Hibou: A family of foundational vision transformers for pathology. *arXiv preprint arXiv:2406.05074*, 2024.

Kutsev Bengisu Ozyoruk 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. 2024.

Pushpak Pati et al. Accelerating histopathology workflows with generative ai-based virtually multiplexed tumour profiling. *Nature Machine Intelligence*, 6(9):1077–1093, 2024.

Qiong Peng et al. Advancing H&E-to-IHC virtual staining with task-specific domain knowledge for HER2 scoring. In *MICCAI*, pp. 3–13. Springer, 2024.

Yue Peng et al. Usigan: Unbalanced self-information feature transport for weakly paired image ihc virtual staining. *IEEE TIP*, 2026.

Elia Peruzzo, Adil Karjauv, Nicu Sebe, Amir Ghodrati, and Amir Habibian. ADAPTOR: Adaptive token reduction for video diffusion transformers. In *CVPR*, pp. 6365–6371, 2025.

Hao Phung, Quan Dao, Trung Dao, Hoang Phan, Dimitris Metaxas, and Anh Tran. DiMSUM: Diffusion mamba–a scalable and unified spatial-frequency method for image generation. *NIPS*, 37: 32947–32979, 2024.

Colin A Purdie et al. HER2 assessment on core biopsy specimens using monoclonal antibody CB11 accurately determines HER2 status in breast carcinoma. *Histopathology*, 56(6):702–707, 2010.

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.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *ECCV*, pp. 87–103. Springer, 2024.

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.

Hoang Thanh-Tung and Truyen Tran. Catastrophic forgetting and mode collapse in GANs. In *IJCNN*, pp. 1–10. IEEE, 2020.

Jiahao Wang et al. LiT: Delving into a simplified linear diffusion transformer for image generation. *arXiv preprint arXiv:2501.12976*, 2025a.

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 *ICIP*, 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 *CVPR*, pp. 8798–8807, 2018.

Tong Wang et al. ODA-GAN: Orthogonal decoupling alignment GAN assisted by weakly-supervised learning for virtual immunohistochemistry staining. In *CVPR*, pp. 25920–25929, 2025b.

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.

John N Weinstein et al. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45 (10):1113–1120, 2013.

Enze Xie et al. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. 2025.

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 *AAAI*, volume 39, pp. 8780–8787, 2025.

Hanwen Xu et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 2024.

Shurong Yang et al. D-VST: Diffusion transformer for pathology-correct tone-controllable cross-dye virtual staining of whole slide images. In *NIPS*, 2025.

Srikar Yellapragada et al. Pixcell: A generative foundation model for digital histopathology images. *arXiv preprint arXiv:2506.05127*, 2025a.

Srikar Yellapragada et al. Pathology image compression with pre-trained autoencoders. In *MICCAI*, pp. 442–452. Springer, 2025b.

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 *ACM MM*, pp. 10210–10219, 2025.

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

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 *MICCAI*, pp. 178–188. Springer, 2024.

Xiaoyu Zheng et al. Diffusion-based virtual staining from polarimetric mueller matrix imaging. In *MICCAI*, pp. 163–173. Springer, 2025.

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

Lianghui Zhu et al. Dig: Scalable and efficient diffusion models with gated linear attention. In *CVPR*, pp. 7664–7674, 2025.

Eric Zimmermann et al. Virchow2: Scaling self-supervised mixed magnification models in pathology. *arXiv preprint arXiv:2408.00738*, 2024.

A Details of the Flow Matching Network

The proposed StainFlash adopts a flow matching network composed of 28 basic blocks as its backbone. Each basic block consists of self-attention, cross-attention, a feedforward layer, and two modulation mechanisms. As illustrated in the left part of Figure S1, the modulation mechanism operates before and after both the self-attention and feedforward layers through element-wise scaling and shifting, which are conditioned on $\mathbf{V}_T$ and $\mathbf{V}_I$. The element-wise sum of $\mathbf{V}_T$ and $\mathbf{V}_I$ is projected into six embeddings: $\boldsymbol{\alpha}_s$, $\boldsymbol{\beta}_s$, $\mathbf{g}_s$, $\boldsymbol{\alpha}_m$, $\boldsymbol{\beta}_m$, and $\mathbf{g}_m$. These embeddings are injected into each basic block. In every basic block, six learnable embeddings—$\boldsymbol{\alpha}_s^i$, $\boldsymbol{\beta}_s^i$, $\mathbf{g}_s^i$, $\boldsymbol{\alpha}_m^i$, $\boldsymbol{\beta}_m^i$, and $\mathbf{g}_m^i$—are element-wise added to the projected embeddings above, producing the final modulation parameters for this block:

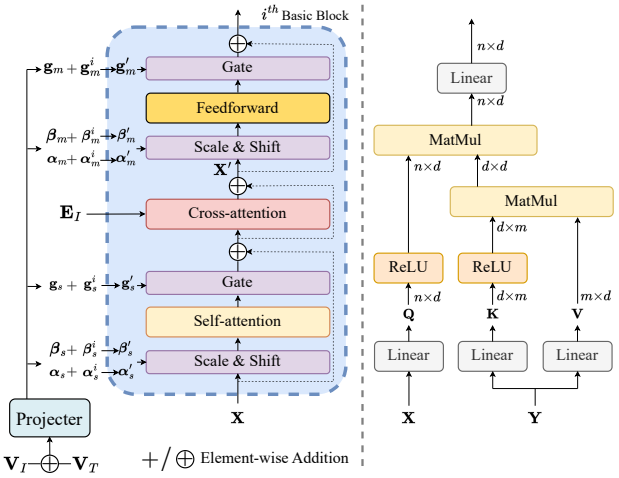

Figure S1: Left: Architecture of a basic block in the flow matching network. Right: linear attention-based cross-attention computation pipeline (it becomes self-attention when $\mathbf{X} = \mathbf{Y}$). In StainFlash, both the self-attention and cross-attention layers are instantiated using linear attention.

$\boldsymbol{\alpha}'_s$, $\boldsymbol{\beta}'_s$, $\mathbf{g}'_s$, $\boldsymbol{\alpha}'_m$, $\boldsymbol{\beta}'_m$, and $\mathbf{g}'_m$. Let the inputs to the self-attention and feedforward layers be $\mathbf{X}$ and $\mathbf{X}'$, respectively. In "Scale & Shift":

$$\mathbf{X} = (1 + \boldsymbol{\alpha}'_s) * \mathrm{LN}(\mathbf{X}) + \boldsymbol{\beta}'_s, \tag{7}$$

$$\mathbf{X}' = (1 + \boldsymbol{\alpha}'_m) * \mathrm{LN}(\mathbf{X}') + \boldsymbol{\beta}'_m, \tag{8}$$

where LN denotes LayerNorm in PyTorch. After $\mathbf{X}$ and $\mathbf{X}'$ passing through the self-attention or feedforward layer, the "Gate" operation is applied as follows:

$$\mathbf{X} = \mathbf{g}'_s * \mathbf{X}, \tag{9}$$

$$\mathbf{X}' = \mathbf{g}'_m * \mathbf{X}', \tag{10}$$

which produces the final output of the self-attention or feedforward module. In our implementation, both the self-attention and cross-attention layers are instantiated using linear attention Katharopoulos et al. (2020) for better efficiency. We also provide an illustration of a linear attention-based cross-attention computation pipeline in the right part of Figure S1 (it becomes self-attention when $\mathbf{X} = \mathbf{Y}$), to support understanding of the textual description in the main text.

B Datasets

We conduct extensive experiments on four H&E-to-IHC virtual staining datasets: RegH2I Peng et al. (2024), BCI Liu et al. (2022a), MIST-HER2 Li et al. (2023), and MIST-ER Li et al. (2023). **RegH2I** Peng et al. (2024) contains 1,992 training and 600 testing pairs. **BCI** Liu et al. (2022a) provides 3,896 training and 977 testing pairs. **MIST-HER2** Li et al. (2023) includes 4,642 training and 1,000 testing pairs. The IHC images in these three datasets are labeled with the HER2 biomarker. **MIST-ER** Li et al. (2023), labeled with the ER biomarker, contributes 4,153 training and 1,000 test pairs. We follow official train/validation/test splits and evaluate on cropped image patches of 1024×1024 pixels Li et al. (2023).

C Potential Tissue Distortion

In virtual staining, paired H&E-IHC images are typically obtained from serial or re-stained tissue sections and algorithmically registered; minor morphological discrepancies inherent to this process mean that alignment is not pixel-perfect. However, most pixels are closely matched, and the task objective—predicting protein expression rather than pixel-level identity—is consistent with routine pathology practice and does not require exact correspondence. Prior methods have successfully

Table S1: Evaluation of various methods on the MIST-Ki67 (left), the MIST-PR (middle), and the HER2match datasets (right).

Method	MIST-Ki67			MIST-PR			HER2match		
	DISTS↓	FID↓	KID↓	DISTS↓	FID↓	KID↓	DISTS↓	FID↓	KID↓
CycleGAN Zhu et al. (2017)	0.241	38.77	0.0122	0.263	45.69	0.0113	**0.187**	24.41	0.0081
AI-FFPE Ozyoruk et al. (2021)	0.246	36.98	0.0090	0.272	49.93	0.0193	0.192	23.78	0.0109
USIGAN Peng et al. (2026)	0.230	27.93	0.0057	0.233	35.15	0.0064	**0.187**	25.68	0.0132
Pix2pix Isola et al. (2017)	0.240	64.99	0.0441	0.271	75.11	0.0420	0.211	33.80	0.0231
Pix2pixHD Wang et al. (2018)	0.244	53.29	0.0331	0.259	60.08	0.0239	0.203	25.15	0.0163
Peng et al. Peng et al. (2024)	0.227	44.95	0.0223	0.246	65.80	0.0323	0.196	24.80	0.0132
ControlNet Zhang et al. (2023)	0.328	250.35	0.2851	0.355	272.90	0.2609	0.414	225.94	0.2640
HistDiST Großkopf et al. (2025)	0.280	73.67	0.0453	0.282	81.35	0.0452	0.229	58.51	0.0607
StainFuser Jewsbury et al. (2024)	0.248	63.15	0.0475	0.256	72.06	0.0377	0.276	107.22	0.0937
pix2pix-Turbo Parmar et al. (2024)	0.384	161.50	0.1529	0.445	78.62	0.0436	0.192	17.65	0.0084
PixCell Yellapragada et al. (2025a)	0.214	42.03	0.0179	0.207	45.28	0.0155	0.229	71.23	0.0621
StainFlash (ours)	**0.196**	**23.79**	**0.0019**	**0.200**	**28.05**	**0.0013**	**0.187**	**16.56**	**0.0077**

Table S2: Performance evaluation of various methods using *pathology-oriented* metrics on RegH2I, RegH2I-Ext, BCI, and MIST-HER2 datasets. C-FID: CONCH-based FID. C-SIM: Mean cosine similarity score computed using CONCH-extracted embeddings.

Method	RegH2I		RegH2I-Ext		BCI		MIST-HER2	
	C-FID↓	C-SIM↑	C-FID↓	C-SIM↑	C-FID↓	C-SIM↑	C-FID↓	C-SIM↑
CycleGAN Zhu et al. (2017)	34.17	0.8904	102.28	0.8546	79.13	0.7586	98.25	0.7614
AI-FFPE Ozyoruk et al. (2021)	30.31	0.8793	98.60	0.8381	130.76	0.6761	93.35	0.7285
USIGAN Peng et al. (2026)	78.91	0.8447	127.80	0.8250	181.84	0.6892	31.14	0.8009
Pix2pix Isola et al. (2017)	61.63	0.8678	227.92	0.7055	232.08	0.6255	194.81	0.6088
Pix2pixHD Wang et al. (2018)	38.49	0.8866	107.79	0.8172	204.55	0.6374	152.63	0.6684
Peng et al. Peng et al. (2024)	23.48	0.9123	125.64	0.7857	155.43	0.6572	85.53	0.7202
ControlNet Zhang et al. (2023)	345.27	0.5995	391.27	0.5740	302.82	0.5931	312.35	0.5946
HistDiST Großkopf et al. (2025)	59.49	0.8520	101.87	0.8177	54.45	0.7784	166.70	0.6604
StainFuser Jewsbury et al. (2024)	101.23	0.8052	138.99	0.7899	112.80	0.6934	74.91	0.7142
pix2pix-Turbo Parmar et al. (2024)	359.92	0.6055	524.48	0.4519	651.58	0.3619	519.64	0.4028
PixCell Yellapragada et al. (2025a)	57.80	0.8642	140.80	0.7898	50.65	0.8163	53.23	0.8169
StainFlash (ours)	**17.53**	**0.9281**	**56.18**	**0.8747**	**23.12**	**0.8397**	**15.89**	**0.8434**

trained on the same datasets used in this work Li et al. (2023); Liu et al. (2022a); Peng et al. (2024). Our strong HER2 classification results (Table 3) and favorable pathologist evaluation further confirm StainFlash's robustness to such data-level imperfections as well as superiority to other methods.

D Additional Experiments

D.1 Quantitative Results on Additional Datasets

We conduct further experiments on three more datasets: **MIST-Ki67** Li et al. (2023), which is labeled with the Ki67 biomarker and contains 4,361 training and 1,000 testing pairs, **MIST-PR** Li et al. (2023), which is labeled with the PR biomarker and contains 4,139 training and 1,000 testing pairs, and **HER2match** Klöckner et al. (2025), which is labeled with the HER2 biomarker and contains 11,610 training and 5,980 testing pairs. The experiments are conducted under the same setting as Table 1 in the main text. Table S1 demonstrates StainFlash's superior performance compared with existing methods, and its generalizability across various metrics and datasets.

D.2 Pathology-Oriented Evaluation Metrics

To provide a more comprehensive assessment of pathological fidelity, we further adopt the pathology-specific vision-language pretrained model CONCH Lu et al. (2024). Specifically, we quantify the domain gap between the generated images and the target domains by computing the mean cosine similarity between feature embeddings extracted by the image encoder of CONCH for each paired sample across the RegH2I, RegH2I-Ext, BCI, MIST-HER2, MIST-ER, MIST-Ki67, MIST-PR, and HER2match datasets. Moreover, following Bhosale et al. (2025), we substitute the Inception network used for FID computation with CONCH, thereby aligning the metric evaluation more closely with

Table S3: Evaluation of computational complexity at the resolutions of 2048×2048 and 4096×4096 pixels on an 80-GB GPU in inference latency and memory consumption.

Method	2048×2048		4096×4096	
	Latency	Memory	Latency	Memory
StainFuser Jewsbury et al. (2024)	38.559s	10.579G	537.54s	13.482G
PixCell Yellapragada et al. (2025a)	20.442s	9.1949G	218.05s	11.580G
StainFlash	**4.1783**s	**8.8354**G	**15.895**s	**9.0507**G

Table S4: Performance evaluation of various methods using *pathology-oriented* metrics on MIST-ER, MIST-Ki67, MIST-PR, and HER2match datasets. C-FID: CONCH-based FID. C-SIM: Mean cosine similarity score computed using CONCH-extracted embeddings.

Method	MIST-ER		MIST-Ki67		MIST-PR		HER2match	
	C-FID↓	C-SIM↑	C-FID↓	C-SIM↑	C-FID↓	C-SIM↑	C-FID↓	C-SIM↑
CycleGAN Zhu et al. (2017)	56.99	0.7585	38.04	0.8118	44.33	0.7473	79.94	0.8325
AI-FFPE Ozyoruk et al. (2021)	134.63	0.6987	60.53	0.7626	74.06	0.7058	52.87	0.8462
USIGAN Peng et al. (2026)	27.87	0.8062	24.22	0.8314	36.06	0.7915	47.34	0.8841
Pix2pix Isola et al. (2017)	157.02	0.6379	162.32	0.6572	202.78	0.5813	48.85	0.8689
Pix2pixHD Wang et al. (2018)	113.56	0.6989	170.21	0.6704	144.88	0.6557	30.93	0.8793
Peng et al. Peng et al. (2024)	107.79	0.7109	111.14	0.7204	117.57	0.6913	42.44	0.8551
ControlNet Zhang et al. (2023)	325.08	0.5709	332.63	0.5687	369.23	0.5520	454.18	0.5109
HistDiST Großkopf et al. (2025)	164.53	0.6591	132.22	0.6948	154.15	0.6609	91.10	0.8748
StainFuser Jewsbury et al. (2024)	91.84	0.7050	105.43	0.6851	118.86	0.6872	244.39	0.6114
pix2pix-Turbo Parmar et al. (2024)	496.49	0.4096	427.89	0.4866	358.06	0.5178	50.61	0.8664
PixCell Yellapragada et al. (2025a)	55.68	0.8206	61.97	0.8159	49.42	0.8250	149.13	0.7439
StainFlash (ours)	**14.90**	**0.8600**	**17.67**	**0.8663**	**16.02**	**0.8481**	**21.38**	**0.8948**

the pathology domain. As shown in Table S2 and Table S4, StainFlash attains the highest CONCH similarity and the lowest FID scores, highlighting its superior ability to produce clinically relevant virtual staining results and corroborating the effectiveness of our design under pathology-oriented evaluation criteria. These findings are consistent with the results observed in our virtual staining and downstream task experiments in the main text.

D.3 Additional Efficiency Analysis

In Table 2 of the main text, we have benchmarked the computational complexity of the proposed StainFlash by measuring the inference latency and memory consumption when generating images of 1024×1024 pixels, where StainFlash demonstrated lower latency and memory use than StainFuser (diffusion-based) Jewsbury et al. (2024) and PixCell (flow matching based) Yellapragada et al. (2025a). Here, we further benchmark the computational complexity when generating images of 2048×2048 and 4096×4096 pixels. Most notably, Table S3 indicates that StainFlash's advantages in both latency and memory become more prominent with increasing resolutions. Compared with StainFuser, StainFlash requires only 10.8% of the latency and 83.5% of the memory for images of 2048×2048 pixels, and 2.9% of the latency and 67.1% of the memory for images of 4096×4096 pixels. Meanwhile, PixCell incurs ∼4.89 times more inference time and ∼1.04 times more memory than StainFlash for images of 2048×2048 pixels, and ∼13.7 times more inference time and ∼1.27 times more memory for images of 4096×4096 pixels. These results highlight the substantial computational advantages of StainFlash.

D.4 Patch Token Sampling Analysis

The proposed KIIF module uses all patch tokens from the pathology foundation models UNI2-h Chen et al. (2024b) and Phikonv2 Filiot et al. (2024). In this section, we further investigate the performance-efficiency trade-off by uniformly sampling a subset of patch tokens. As shown in Table S5, uniformly sampling 25% or 50% of the patch tokens yields slightly degraded performance with marginal efficiency improvements. Given the limited efficiency gains, we use all patch tokens in StainFlash and leave more sophisticated and more effective sampling strategies for future exploration.

Table S5: Performance-efficiency trade-off study on the RegH2I dataset concerning the ratio of patch tokens used in the KIIF module.

Patch token ratio	DISTS↓	FID↓	KID↓	Latency	Memory
25%	0.175	32.45	0.0068	1.129s	8.764G
50%	0.174	31.96	0.0061	1.140s	8.769G
100% (StainFlash)	**0.173**	**31.66**	**0.0033**	**1.149**s	**8.781**G

Table S6: Performance comparison on the RegH2I dataset of combining **two** of five pathology foundation models.

Foundation models	DISTS↓	FID↓	KID↓
UNI + Prov	0.1741	32.598	0.0064
UNI + Mid	0.1740	33.611	0.0054
UNI + Vir	0.1749	32.997	0.0080
Phi + Prov	0.1743	33.362	0.0062
Phi + Mid	0.1748	33.236	0.0049
Phi + Vir	0.1761	32.975	0.0068
Prov + Mid	0.1758	33.954	0.0077
Prov + Vir	0.1739	32.607	0.0069
Mid + Vir	0.1769	34.535	0.0100
UNI + Phi (StainFlash)	**0.1738**	**31.661**	**0.0033**

D.5 Ablation Study on Pathology Foundation Models

StainFlash leverages two representative and powerful foundation models, UNI2-h Chen et al. (2024b) and Phikonv2 Filiot et al. (2024), to extract complementary representations. To explore the design space of various pathology foundation models, we additionally consider three models: Virchow2 Zimmermann et al. (2024) trained on 3.1M whole-slide images (WSIs), Midnight-12k Karasikov et al. (2025) trained on the publicly available TCGA Weinstein et al. (2013) dataset (12k WSIs), and Prov-GigaPath Xu et al. (2024) trained on 1.3 billion pathology tiles of 256×256 pixels from 171,189 whole slides. For convenience, we abbreviate the models' names as follows: "Vir" for "Virchow2", "Prov" for "Prov-GigaPath", "Mid" for "Midnight-12k", "UNI" for "UNI2-h", and "Phi" for "Phikonv2". We first evaluate all possible combinations of two models. As shown in Table S6, most combinations achieve competitive performance relative to the leading methods in Table 1 of the main text. These results suggest that StainFlash can also effectively incorporate other pathology foundation models for consistently strong performance.

We then explore incorporating more pathology foundation models beyond two. As we observe from Table S6 that the best two-model configuration is UNI2-h and Phikonv2 (which is actually the proposed StainFlash), we add additional model(s) on top of them (Table S7). Three-model combinations consistently yield marginally better DISTSs, but worse FIDs and KIDs, with the best configuration being UNI2-h + Phikonv2 + Virchow2. On top of that, we further experiment with combining four and five models. As we can see, most of these configurations produce better DISTSs and FIDs than StainFlash but worse KIDs, though the differences are again marginal. This is likely because additional models provide redundant information to the representations jointly extracted by UNI2-h and Phikonv2 within the current StainFlash framework. Given that incorporating more models inevitably increases system complexity and inference latency, we adopt the two-model configuration with only UNI2-h and Phikonv2 in StainFlash.

D.6 Potential Impact of Downsampling Ratio on Quality

We compare the IHC image reconstruction quality of autoencoders with different downsampling rates, including: 1) the 8× downsampling variational autoencoder (VAE) provided by Stable Diffusion 3 Esser et al. (2024) and adopted in PixCell Yellapragada et al. (2025a), and 2) the 32× downsampling DC-AE provided by SANA Xie et al. (2025) and adopted by our StainFlash, on the RegH2I dataset. Table S8 shows that the 8× VAE achieves consistently better metrics than the 32× DC-AE. This is reasonable as the former retains more details for image reconstruction. Yet, StainFlash

Table S7: Performance comparison on the RegH2I dataset of combining **three to five** pathology foundation models.

Foundation models	DISTS↓	FID↓	KID↓
UNI + Phi + Vir	0.1714	31.824	0.0042
UNI + Phi + Prov	0.1717	32.417	0.0044
UNI + Phi + Mid	0.1711	32.598	0.0044
UNI + Phi + Vir + Prov	0.1735	31.440	0.0041
UNI + Phi + Vir + Mid	0.1752	31.401	0.0057
UNI + Phi + Vir + Prov + Mid	**0.1709**	**31.373**	**0.0039**

Table S8: Comparison of image reconstruction quality of 1) the VAE in Stable Diffusion 3 Esser et al. (2024), and 2) the DC-AE in SANA Xie et al. (2025), on the RegH2I dataset.

Autoencoder	Adopted in	Downsample	DISTS↓	FID↓	KID↓
VAE Esser et al. (2024)	PixCell Yellapragada et al. (2025a)	8×	0.0553	5.4359	0.0004
DC-AE Xie et al. (2025)	StainFlash	32×	0.0697	17.678	0.0026

is well balanced under 32× downsampling, simultaneously achieving superior performance and computational efficiency over existing SOTA diffusion/flow matching-based methods. The advantages in both quality and efficiency demonstrate the effectiveness of StainFlash's design.

D.7 Impact of Pretrained Weights

StainFlash adopts pretrained SANA Xie et al. (2025) as its backbone, following common practice in diffusion and flow matching frameworks that leverage models pretrained on large-scale natural image datasets for improved initialization and performance. In this section, we further evaluate two alternative training strategies: (i) training from scratch without pretrained weights, and (ii) freezing the pretrained SANA weights while fine-tuning the remaining components, with all other settings kept identical to Table 1 in the main text. As shown in Table S9, training from random initialization yields significantly degraded performance, which is expected given the limited number of virtual-staining samples available to train a flow matching network from scratch. Moreover, freezing the pretrained SANA weights while only fine-tuning the other modules also leads to substantially inferior results. These findings indicate that although SANA pretraining provides a strong initialization, the overall performance gain primarily stems from our effective framework design rather than pretraining alone.

D.8 Effects of Fine-Tuning Autoencoder

StainFlash inherits the deep compression autoencoder (DC-AE) Chen et al. (2025b) from SANA Xie et al. (2025), as it adopts SANA's flow matching network as the backbone network. Our empirical evidence, both quantitative and qualitative, shows that StainFlash with DC-AE outperforms various GAN- and diffusion-based approaches on both virtual staining and downstream classification tasks (cf. Table 1, Figure 2, and Table 3 of the main text). Thus, we argue that DC-AE suffices for encoding and decoding of pathology images in StainFlash, despite being pretrained on natural images and probably not being the optimal choice. We attribute this to DC-AE's strong encoding/decoding capabilities from large-scale, diverse training, and to the flow matching network's high compatibility with it. Meanwhile, several recent works, e.g., HistDiST Großkopf et al. (2025), StainFuser Jewsbury et al. (2024), and PixCell Yellapragada et al. (2025a), have also successfully employed variational autoencoders (VAEs) pretrained on natural images to generate histopathology images.

We further experiment with fine-tuning the DC-AE following the practice in Yellapragada et al. (2025b) on the RegH2I training set, incorporating an L1 distance between the embeddings produced by the UNI Chen et al. (2024b) encoder, an L1 pixel reconstruction loss, and a learned PatchGAN discriminator loss Isola et al. (2017). To maintain compatibility with the latent space of the flow matching network in StainFlash, we only fine-tune the decoder of DC-AE. We then evaluate the fine-tuned DC-AE on the RegH2I test set. Table S10 shows that fine-tuning the DC-AE brings marginal improvements in DISTS and FID, while yielding a worse KID. We speculate that the limited fine-tuning data prevents substantial improvements in virtual staining quality, though further verifi-

Table S9: Performance comparison on the RegH2I dataset of not using pretrained weights (w/o pretrain), freezing the pretrained weights (freeze pretrain), and tuning all parameters (StainFlash) for training.

Methods	DISTS↓	FID↓	KID↓
w/o pretrain	0.188	44.66	0.0185
freeze pretrain	0.183	37.32	0.0099
StainFlash	**0.173**	**31.66**	**0.0033**

Table S10: Effects of fine-tuning the DC-AE on the RegH2I dataset.

DC-AE	DISTS↓	FID↓	KID↓
Original Xie et al. (2025)	0.174	31.66	**0.0033**
Fine-tuned	**0.173**	**31.49**	0.0070

cation with larger datasets is necessary. Given that fine-tuning DC-AE adds additional engineering complexity yet yields marginal improvements, we decide to adopt the original SANA DC-AE for experiments in the main text. To encourage further community investigation in this direction, we will release the fine-tuned DC-AE weights along with our code.

D.9 Analysis of the Modulation Factors

StainFlash projects the global timestep T into $\mathbf{V}'_T$, and further map $\mathbf{V}'_T$ to the channel-wise scale α and bias β, which are used to modulate $\mathbf{C}_I$ for timestep-dependent H&E feature injection. In this section, we visualize the norms of α and β as functions of the global timestep T. As shown in Figure S2, the norm of α is relatively large during early timesteps (note that the timesteps gradually decrease as inference proceeds), since abundant structural cues from $\mathbf{C}_I$ are injected into the flow matching network at the early stage of the flow matching process. In the middle and late stages, the norm of α decreases and gradually stabilizes, as the injection becomes more conservative to facilitate fine-grained detail refinement conditioned on H&E images. Meanwhile, the norm of β remains relatively stable throughout the process and is consistently non-zero, indicating that β serves as an adjustment term across all timesteps.

D.10 Analysis of $\mathbf{V}_I$

The KIIF module projects the global tokens from the two models into $\mathbf{V}_I$. The element-wise sum of $\mathbf{V}_I$ and $\mathbf{V}_T$ is then used to generate the modulation factors for each block of the flow matching network. In this section, we analyze the ratio of norms between $\mathbf{V}_I$ and $\mathbf{V}_T$ across different cases to investigate whether $\mathbf{V}_I$ exerts effective control. Specifically, we compute the per-case average norm of $\mathbf{V}_T$ over all global timesteps T of the entire flow matching inference process, denoted by $\|\mathbf{V}_{Tm}\|$. We also compute the norm of $\mathbf{V}_I$ for all cases in the RegH2I test set and draw the histogram of $\|\mathbf{V}_I\|/\|\mathbf{V}_{Tm}\|$ in Figure S3. As we can see, the norm ratios vary across cases. Combined with the ablation results in Table 4 of the main text, this demonstrates that global tokens, injected through $\mathbf{V}_I$, provide tangible performance gains for the flow matching network.

D.11 Failure Analysis

We further analyze a representative failure case of StainFlash (Figure S4). Here, a HER2 2+ region is mistakenly stained as 0, and the moderate membrane staining intensity that defines this intermediate grade is lost. This observation is consistent with the downstream classification analysis in the main text, where the largest performance gaps are observed for the moderate-positive grades (1+ and 2+). 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 than the clearly negative (0) and strongly positive (3+) grades, which we regard as a promising direction for future work.

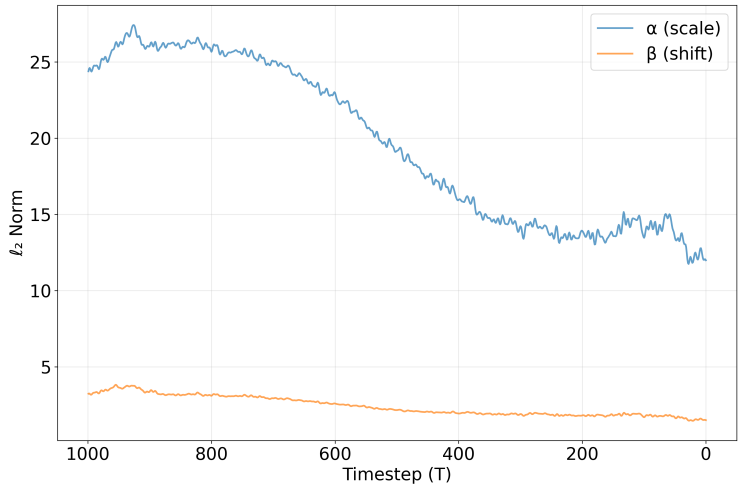

Figure S2: Variations of the norms of α and β with respect to the timestep T.

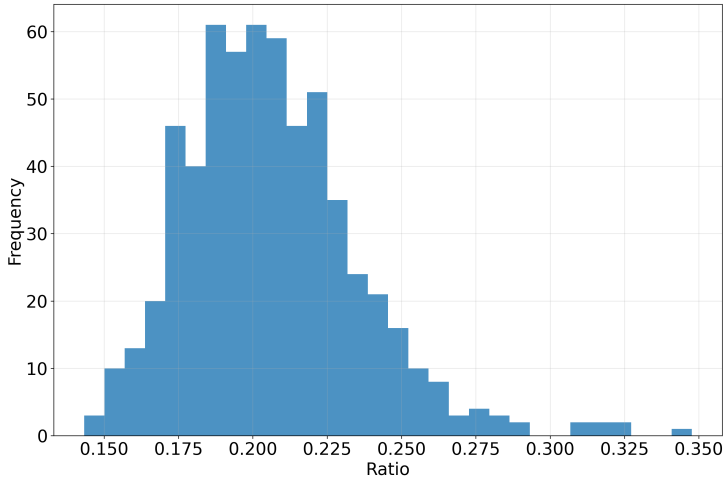

Figure S3: Histogram of the norm ratio between $\mathbf{V}_I$ and $\mathbf{V}_{Tm}$ of the RegH2I test set.

D.12 Additional Qualitative Results

We present additional qualitative results across all datasets, RegH2I, BCI, MIST-HER2, MIST-Ki67, MIST-ER, MIST-PR, and HER2match in Figures S5–S11. As shown, StainFlash achieves higher fidelity, providing a more accurate representation of both structural characteristics and pathological status, as well as superior perceptual realism and visual plausibility compared with other methods.

D.13 Additional WSI results

To further evaluate the generalizability of StainFlash on whole-slide images (WSIs), we additionally test it on three H&E WSIs. These WSIs are labeled with the HER2 biomarker and serve as an external test set. The compared methods are selected based on their quantitative and qualitative performance on the RegH2I dataset, as reported in Table 1 of the main text. Specifically, we include the higher-performance approaches Pix2pixHD Wang et al. (2018), AI-FFPE Ozyoruk et al. (2021), and Peng et al. (2024). For inference on the H&E WSIs, all methods (including ours) use models trained on the RegH2I dataset.

Since WSIs are ultra-high-resolution and full-resolution inference is infeasible, we partition each WSI into partially overlapping patches of 1024×1024 pixels. We perform inference patch-wise, then

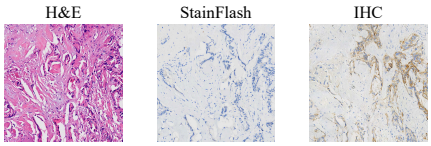

Figure S4: A representative failure case of StainFlash. A HER2 2+ region is mistakenly stained as 0, missing the moderate membrane staining intensity that characterizes the intermediate grade.

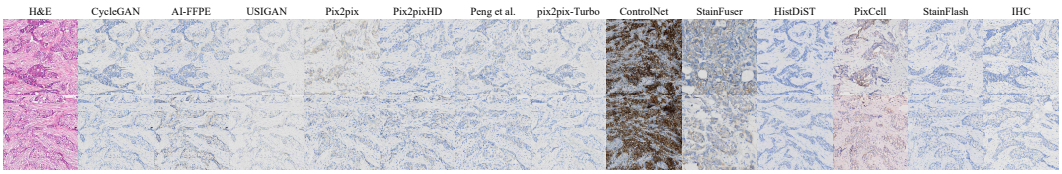

Figure S5: Example virtual staining results on the RegH2I dataset (best viewed zoomed-in). In IHC images, blue dots denote cell nuclei, and brown regions in the cell membranes correspond to cancerous lesions, with darker brown coloration signifying more advanced lesion severity and further disease progression.

reconstruct the full WSI by stitching the outputs with overlaps, following MultiDiffusion Bar-Tal et al. (2023). Blind rankings are collected from two board-certified pathologists, who score the virtual staining results from 1 (best) to 4 (worst), considering both perceptual quality and pathological fidelity of the generated images.

The virtually stained WSIs are shown in Figures S12–S14. The mean ranking of the four methods is: StainFlash (1.0), Peng et al. (2024) (2.333), AI-FFPE (2.667), and Pix2pixHD (3), indicating StainFlash's superior scalability and virtual staining quality. According to the pathologists' feedback, StainFlash produces the highest quality and fidelity of virtual H&E to IHC staining among all compared methods. In particular, StainFlash achieves high perceptual fidelity with staining colors and texture closely aligned with real IHC images, while preserving critical pathological information from the input H&E images. In contrast, the compared methods exhibit inferior staining quality with issues such as overly sharpened colors and textures, blurred tumor cell boundaries, and incorrect cell pathology statuses—factors that significantly impede reliable pathological assessment.

E Limitations and Future Work

StainFlash enables efficient virtual staining via a DC-AE and linear-attention-based flow matching framework. Its inference speed remains constrained by the multi-step denoising process of flow matching Liu et al. (2022b). We plan to optimize the denoising scheduler and reduce inference steps via consistency models Chen et al. (2025a), as well as minimize computational overhead and accelerate inference through model pruning and distillation retraining.

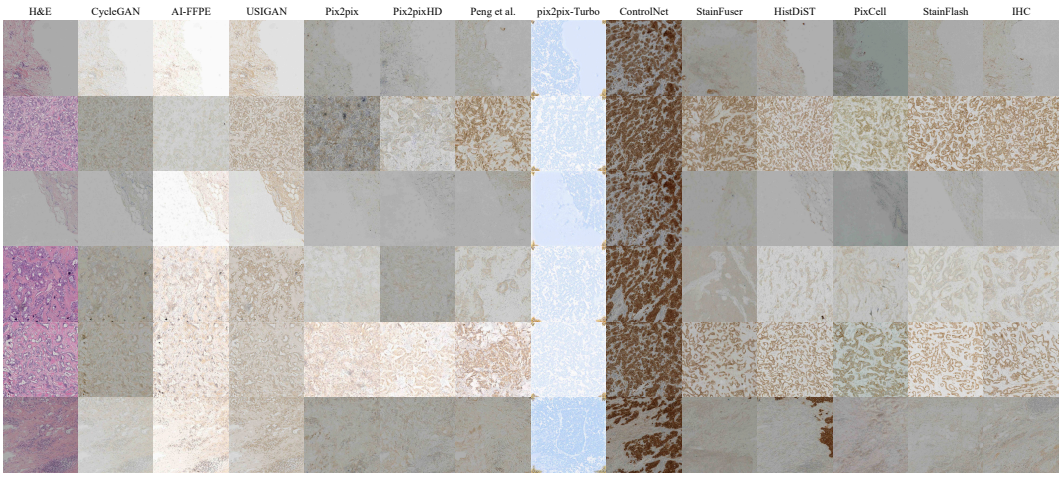

Figure S6: Example virtual staining results on the BCI dataset (best viewed zoomed-in). In IHC images, blue dots denote cell nuclei, and brown regions in the cell membranes correspond to cancerous lesions, with darker brown coloration signifying more advanced lesion severity and further disease progression.

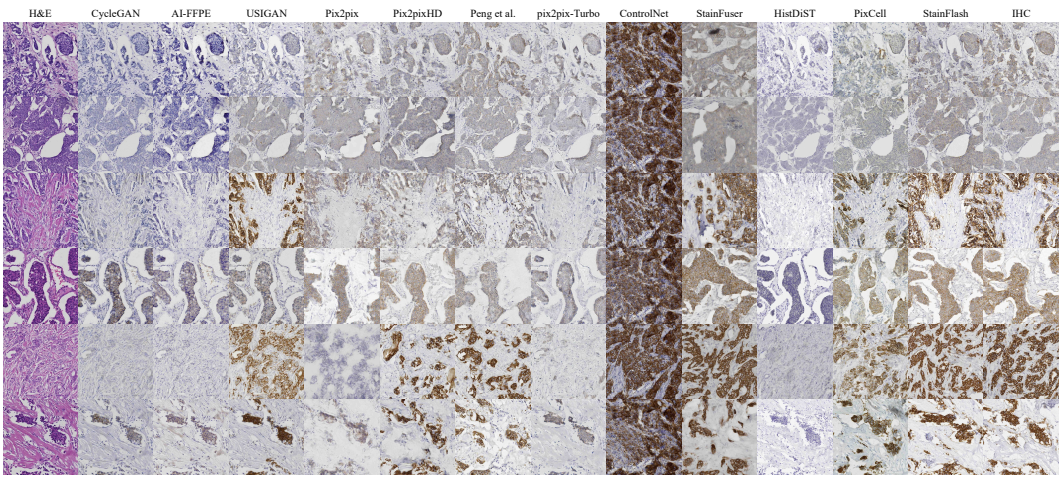

Figure S7: Example virtual staining results on the MIST-HER2 dataset (best viewed zoomed-in). In IHC images, blue dots denote cell nuclei, and brown regions in the cell membranes correspond to cancerous lesions, with darker brown coloration signifying more advanced lesion severity and further disease progression.

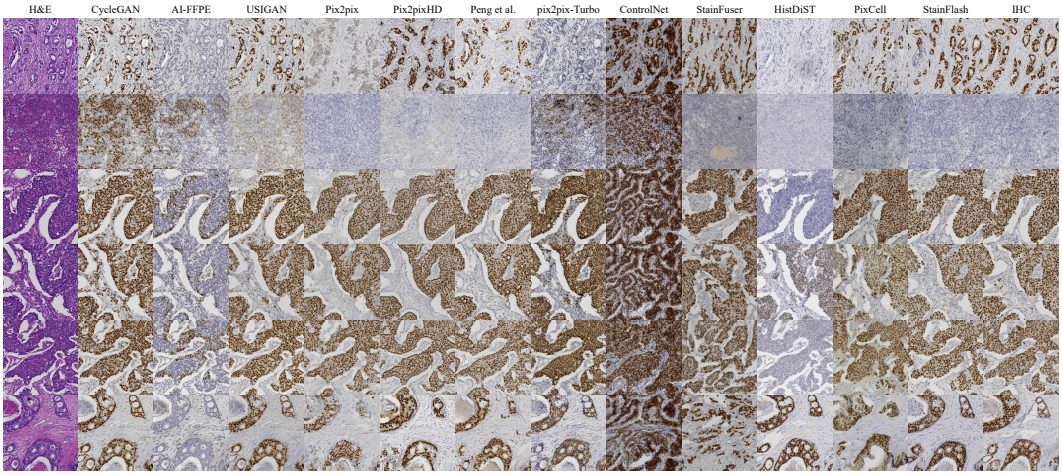

Figure S8: Example virtual staining results on the MIST-ER dataset (best viewed zoomed-in). In IHC images, blue dots represent nuclei of cells that are non-estrogen receptor (ER) expression, while brown dots represent nuclei of receptor expression, with darker brown coloration indicating stronger receptor expression, generally corresponding to higher differentiation and more favorable prognosis.

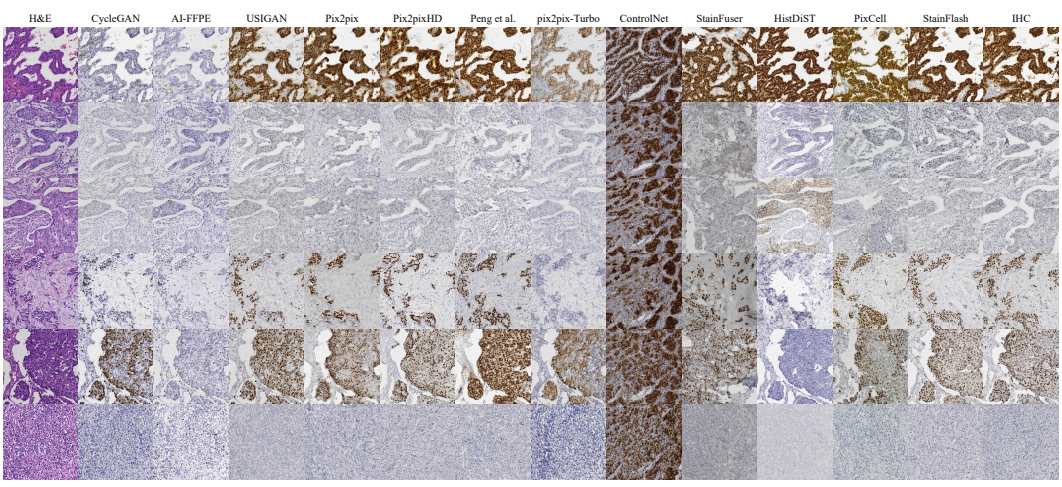

Figure S9: Example virtual staining results on the MIST-PR dataset (best viewed zoomed-in). In IHC images, blue dots represent nuclei of cells that are non-progesterone receptor (PR) expression, while brown dots represent nuclei of receptor expression, with darker brown coloration signifying stronger receptor expression, generally corresponding to higher differentiation and more favorable prognosis.

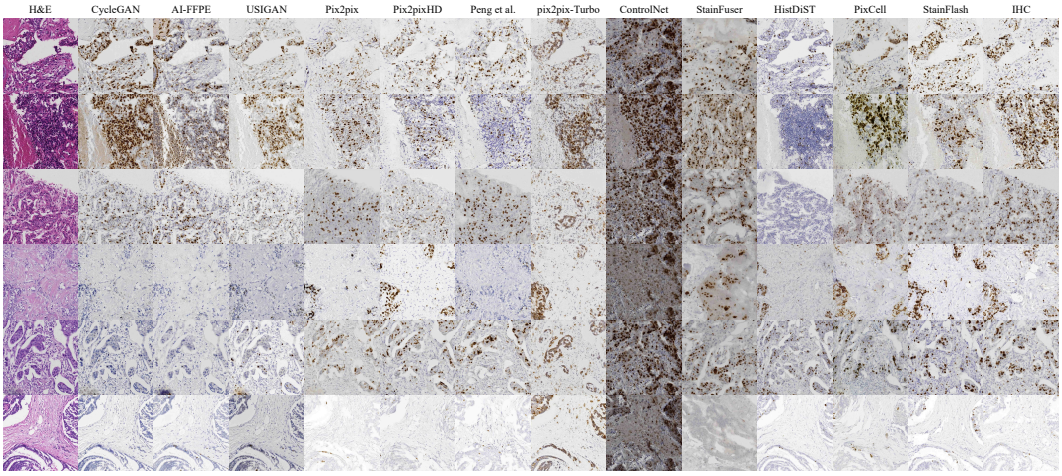

Figure S10: Example virtual staining results on the MIST-Ki67 dataset (best viewed zoomed-in). In IHC images, blue dots represent nuclei of non-proliferating cells, while brown dots indicate nuclei of actively proliferating cells, with a higher density of brown nuclei reflecting increased proliferation, typically associated with more advanced lesion severity and disease progression.

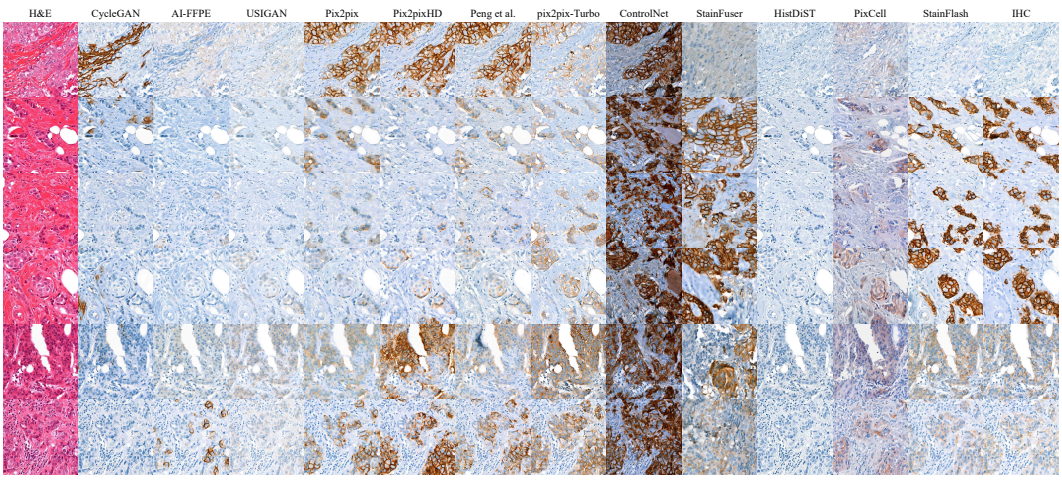

Figure S11: Example virtual staining results on the HER2match dataset (best viewed zoomed-in). In IHC images, blue dots denote cell nuclei, and brown regions in the cell membranes correspond to cancerous lesions, with darker brown coloration signifying more advanced lesion severity and further disease progression.

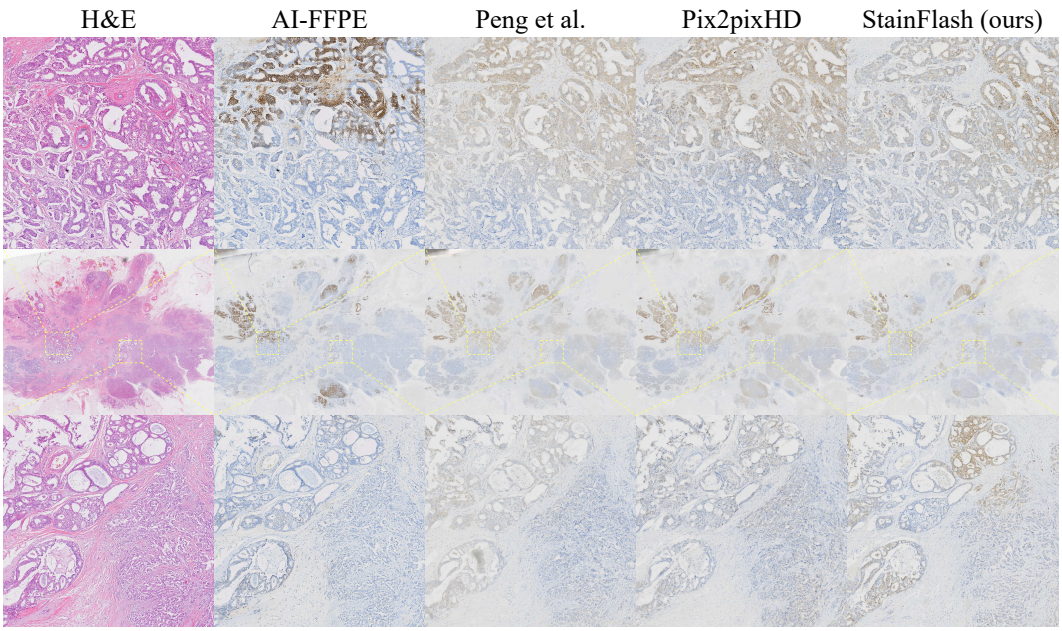

Figure S12: Virtual IHC WSIs generated from an H&E WSI image (resolution: 38144×29984 pixels). Center: WSI images; and top and bottom: zoomed views of two patches of 4096×4096 pixels.

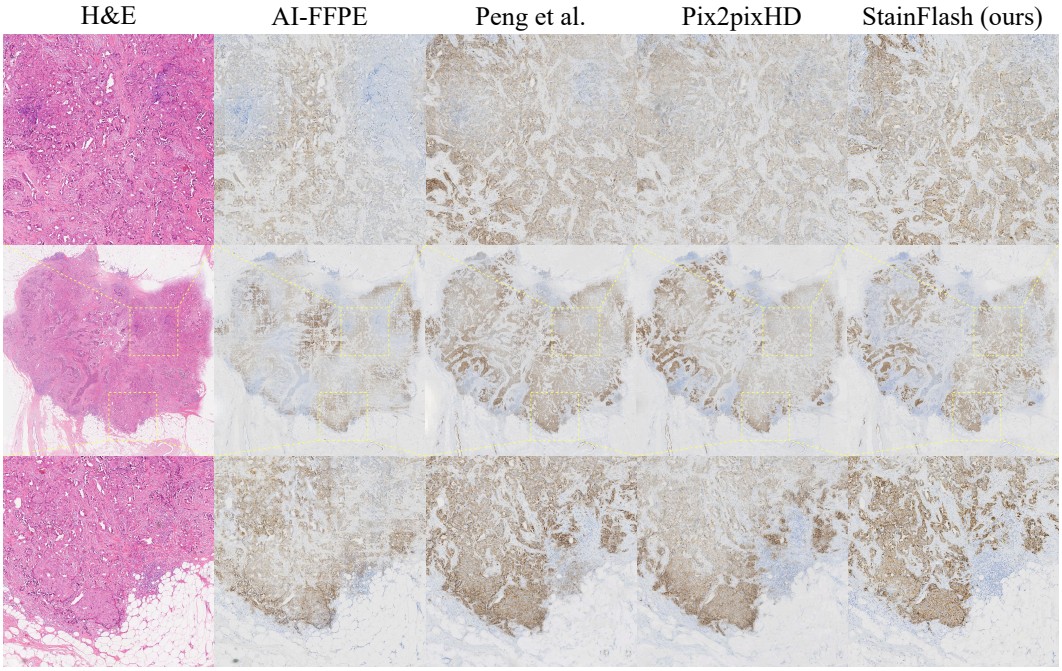

Figure S13: Virtual IHC WSIs generated from an H&E WSI image (resolution: 17984×17984 pixels). Center: WSI images; and top and bottom: zoomed views of two patches of 4096×4096 pixels.

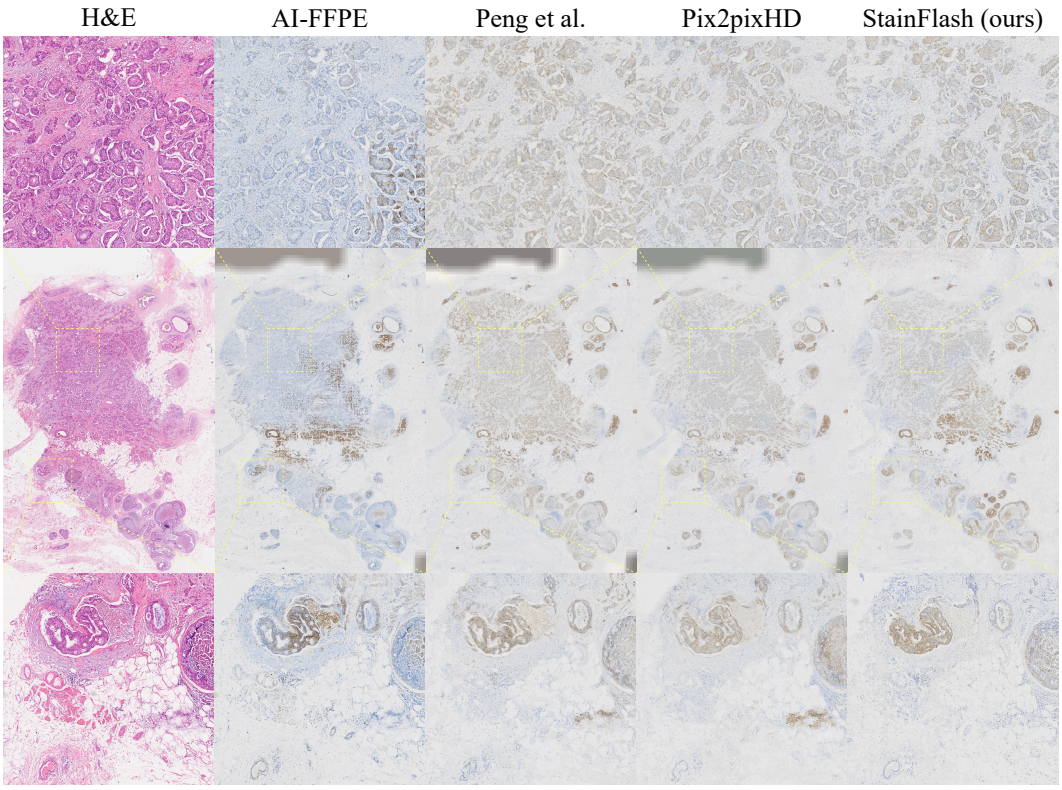

Figure S14: Virtual IHC WSIs generated from an H&E WSI image (resolution: 19872×30720 pixels). Center: WSI images; and top and bottom: zoomed views of two patches of 4096×4096 pixels.