Virtual Multiplex Staining for Histological Images Using a Marker-wise Conditioned Diffusion Model

Hyun-Jic Oh[1], Junsik Kim[2], Zhiyi Shi[2], Yichen Wu[2], Yu-An Chen[3], Peter K Sorger[3], Hanspeter Pfister[2], Won-Ki Jeong[1]*

[1] Korea University
[2] Harvard University
[3] Harvard Medical School, Harvard University
{hyunjic0127,wkjeong}@korea.ac.kr, mibastro@gmail.com, zhiyis@seas.harvard.edu, wuyichen.am97@gmail.com, pfister@seas.harvard.edu, yu-an_chen@hms.harvard.edu, peter_sorger@hms.harvard.edu

Abstract

Multiplex imaging is revolutionizing pathology by enabling the simultaneous visualization of multiple biomarkers within tissue samples, providing molecular-level insights that traditional hematoxylin and eosin (H&E) staining cannot provide. However, the complexity and cost of multiplex data acquisition have hindered its widespread adoption. Additionally, most existing large repositories of H&E images lack corresponding multiplex images, limiting opportunities for multimodal analysis. To address these challenges, we leverage recent advances in latent diffusion models (LDMs), which excel at modeling complex data distributions by utilizing their powerful priors for fine-tuning to a target domain. In this paper, we introduce a novel framework for virtual multiplex staining that utilizes pretrained LDM parameters to generate multiplex images from H&E images using a conditional diffusion model. Our approach enables marker-by-marker generation by conditioning the diffusion model on each marker, while sharing the same architecture across all markers. To tackle the challenge of varying pixel value distributions across different marker stains and to improve inference speed, we fine-tune the model for single-step sampling, enhancing both color contrast fidelity and inference efficiency through pixel-level loss functions. We validate our framework on two publicly available datasets, notably demonstrating its effectiveness in generating up to 18 different marker types with improved accuracy, a substantial increase over the 2-3 marker types achieved in previous approaches. This validation highlights the potential of our framework, pioneering virtual multiplex staining. Finally, this paper bridges the gap between H&E and multiplex imaging, potentially enabling retrospective studies and large-scale analyses of existing H&E image repositories.

1 Introduction

Histopathological analysis is crucial for disease diagnosis and biomedical research, enabling the identification of pathological changes and guiding treatment decisions. Among available techniques, hematoxylin and eosin (H&E) staining has long been the gold standard, often complemented by immunohistochemistry (IHC). Despite their widespread use, both H&E and IHC have limitations in capturing the complex tissue microenvironment, especially in advanced cancer studies (Wharton Jr et al. 2021). To address these limitations, multiplex immunofluorescence (mIF) and multiplex immunohistochemistry (mIHC) imaging have emerged, enabling visualization of multiple biomarkers within a single tissue sample and providing more comprehensive insights into the tissue microenvironment (see Fig. 1) (Sood et al. 2016; Tan et al. 2020; Lin et al. 2023). However, widespread adoption of multiplex imaging is hindered by complex imaging protocols and high staining costs.

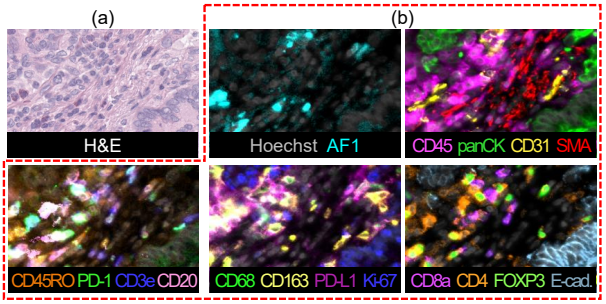

Figure 1: Sample from Orion-CRC (Lin et al. 2023): (a) H&E staining and (b) multiplex immunofluorescence (mIF) of the same region. Each color in mIF corresponds to a different marker.

To overcome these challenges, computational approaches known as *virtual staining*, which generate various staining images from H&E stained tissue images, have gained much attention (Latonen et al. 2024). Most prior work applied deep generative models to unpaired H&E-to-IHC tasks, focusing on spatial alignment (Li et al. 2023; Chen et al. 2024; Wang et al. 2025); other studies explored multiplex marker generation, such as H&E-to-IF (Burlingame et al. 2020) and H&E-to-mIHC (Pati et al. 2024). However, these approaches generally require a separate model per marker, limiting scalability and increasing computational cost, which hinders efficient handling of large marker sets. Moreover, their separate training manner focuses only on individual markers, missing valuable cross-channel relationships and overlooking the potential benefits of knowledge sharing through joint

*Corresponding author: wkjeong@korea.ac.kr

training. Even recent methods like VIMs (Dubey et al. 2024) and HEMIT (Bian et al. 2024), which generate up to two or three markers, still lack the scalability needed for practical multiplex imaging applications.

Based on these observations, we propose a novel framework for virtual multiplex staining that utilizes a conditional diffusion model (Ho, Jain, and Abbeel 2020; Dhariwal and Nichol 2021; Rombach et al. 2022) to translate H&E images into multiplex marker images. By designing a marker-wise generation approach, we aim to generate multiple markers within a single shared model architecture, enabling knowledge sharing during training and reducing the significant computational overhead. Specifically, to optimize both the efficiency of the generation process and the quality of the generated results, we adopt a two-stage training approach, each addressing a core challenge:

(1) Multi-target generation. Unlike previous methods training separate models for each marker or lacking scalability for the large number of marker types, we use a marker-wise conditioning strategy to generate multiple markers simultaneously using a single model. By conditioning H&E images with different marker-type embeddings using latent diffusion models (Rombach et al. 2022), we enable the model to effectively distinguish among a large set of markers and generate high-quality multiplex staining from H&E images, overcoming the limitations of conventional text-based conditioning when the number of marker types increases. Therefore, the training strategy reduces computational costs and enables potential knowledge sharing.

(2) Color contrast fidelity. Due to variations in pixel value ranges and contrasts across marker images, simply applying diffusion models tends to be influenced by the training data bias, where a substantial portion of images contain dark background pixels. To address this issue, we incorporate a second training stage that fine-tunes the model using pixel-level loss functions to enhance marker-specific generation capabilities, addressing color accuracy. This stage also optimizes the model for fast, single-step inference, enabling practical deployment in pathology where timely analysis impacts clinical workflows and large-scale research studies.

Through the proposed two-stage training framework, we efficiently generate high-quality multiplex marker images from H&E staining, preserving spatial and structural context and improving color fidelity, thereby promoting the widespread adoption of multiplex imaging. To the best of our knowledge, this is the first work to demonstrate virtual staining of up to 18 distinct marker images from a single H&E image using a conditional diffusion model. To summarize, our contributions are as follows:

- We propose a novel two-stage training framework for virtual multiplex staining from H&E tissue sections, leveraging a specially designed conditional diffusion model embedded with marker-type embeddings to effectively handle a large number of marker types.
- We address the issue of color fidelity during the second-stage fine-tuning phase by integrating pixel-level loss functions, while also opting to optimize the single-step sampling of the conditional diffusion model for faster inference.
- We validate our framework on two publicly available datasets, demonstrating its effectiveness in generating up to 18 distinct marker types with superior accuracy.

2 Related Work

2.1 Multiplex Imaging for Pathology

H&E staining is the gold standard for tissue examination in pathology, with IHC staining complementing morphological assessment by providing additional molecular information. Previous virtual staining studies have mainly focused on translating H&E to IHC images in unpaired settings to address spatial alignment challenges inherent between these stains (Li et al. 2023; Chen et al. 2024; Wang et al. 2025). While these methods have advanced unpaired translation, their objectives and data settings differ from our focus on paired H&E-to-multiplex virtual staining. Moreover, both H&E and IHC stainings are limited in capturing the complex tissue microenvironment, particularly in advanced cancers (Wharton Jr et al. 2021). This has led to the increased interest in multiplex imaging techniques, which simultaneously visualize multiple biomarkers within a single tissue section—enhancing cellular, molecular, and spatial understanding of pathology (Tan et al. 2020; Wharton Jr et al. 2021; Lin et al. 2023). Despite its analytic advantages, the widespread adoption of multiplex imaging is hindered by resource-intensive protocols and limited data availability.

To address these challenges, research has shifted toward deep learning-based virtual staining methods that generate multiplex images from H&E sections (Burlingame et al. 2020; Pati et al. 2024; Bian et al. 2024). Early efforts relied on pix2pix (Isola et al. 2017), a conditional GAN used as a baseline for paired image-to-image translation from H&E to specific immunomarkers, which required a separate model per marker, thus limiting scalability as the number of biomarkers increases (Burlingame et al. 2020). Recently, HEMIT (Bian et al. 2024) combined residual CNNs with Swin Transformers to jointly generate up to three mIHC markers, partially capturing cross-marker relationships; however, it has yet to demonstrate scalability to larger marker panels. Meanwhile, VIMs (Dubey et al. 2024) used an unpaired, diffusion-based approach to generate two IHC markers with performance comparable to pix2pix, but it requires expert-designed text prompts for marker control, which poses barriers when scaling to more markers or deploying in automated pipelines.

To sum up, existing approaches remain limited by either per-marker model requirements, restricted marker panel size, or dependencies on manual prompt engineering. Efficient joint modeling and scalability to high marker counts, as well as comprehensive knowledge sharing, thus remain open challenges. Our approach directly addresses these limitations by introducing a conditional diffusion model-based framework capable of generating multiple marker types, paving the way for more scalable and practical virtual multiplex imaging from H&E staining.

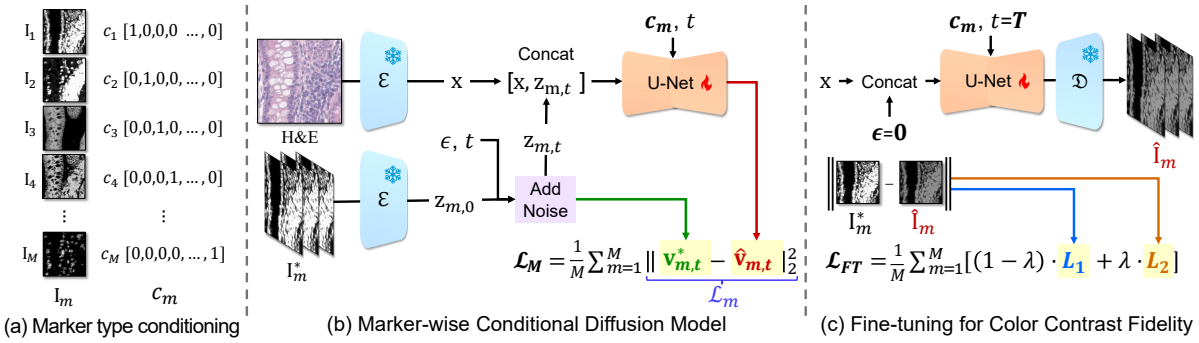

Figure 2: Overview of the proposed two-stage training framework for virtual multiplex staining. (a) Marker type conditioning using the one-hot vector c_m for marker m. (b) Marker-wise conditional diffusion model that generates multiplex marker images from an H&E input with marker-specific one-hot conditioning. (c) Pixel-level fine-tuning stage to improve color fidelity, enabling single-step inference.

2.2 Conditional Diffusion Models

Diffusion models (Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020) have emerged as a powerful generative framework, achieving remarkable fidelity in image synthesis via iterative denoising. Of particular interest, Latent Diffusion Models (LDMs) (Rombach et al. 2022) leverage a Variational Autoencoder (VAE) trained on large-scale datasets to operate in a compressed latent space, substantially improving computational efficiency and providing strong priors for rapid adaptation in diverse domains.

Recent studies have demonstrated the versatility of diffusion models for a wide range of image generation. Conditional diffusion models allow fine-grained control over image generation by leveraging diverse conditioning inputs, including text (Saharia et al. 2022), semantic maps (Zhang, Rao, and Agrawala 2023), or images (Tumanyan et al. 2023). Notably, these models have shown strong performance in dense prediction tasks like semantic segmentation, depth estimation, and structured image-to-image translation by effectively integrating conditioning signals into the diffusion process (Zhao et al. 2023; Ke et al. 2024; Fu et al. 2024). Building on these developments, approaches like Marigold (Ke et al. 2024) keep the VAE prior from pre-trained LDMs and fine-tune only the denoising U-Net for specific tasks, such as depth prediction, achieving efficient adaptation with reduced training complexity. Similarly, Geowizard (Fu et al. 2024) introduces a class-wise one-hot conditioning strategy to differentiate multiple geometric and domain classes, which is beneficial when scaling to many output targets conditioned on the same image prompt.

Despite these advances, existing conditional diffusion models remain under-explored for multiplexed pathology images, which pose distinct challenges such as channel scalability, color fidelity, and computational efficiency. Our work addresses these challenges by leveraging robust priors from pre-trained LDMs and introducing marker-wise conditioning for scalable, high-fidelity multiplex virtual staining.

3 Method

In this section, we introduce our two-stage training framework for virtual multiplex staining in H&E images using a conditional diffusion model for marker-specific generation. The two stages address key challenges of multiplexed image generation and color contrast fidelity using marker specific conditioning as illustrated in Fig. 2(a), with further details provided in Sec. 3.1 for (b) a marker-wise conditional diffusion model and in Sec. 3.2 for (c) a fine-tuning process to enhance color fidelity.

3.1 Marker-wise Conditional Diffusion Model

For the multiplex image generation, we build our virtual staining model on the pre-trained stable diffusion (SD) backbone (Rombach et al. 2022), leveraging its LDM architecture. Specifically, we formulate the virtual staining as a conditional diffusion process that maps H&E images to generate multiplex marker images.

Image encoding. We replicate a single-channel marker image $\mathbf{I}_m$ into three channels to match the input shape of the pre-trained SD VAE $\mathcal{E}$, where $m \in \{1, 2, ..., M\}$ denotes a marker type. During training, we encode the ground-truth marker image through the VAE encoder to obtain the clean latent representation $\mathbf{z}_{m,0} = \mathcal{E}(\mathbf{I}_m)$. Correspondingly, the H&E image is encoded as $\mathbf{x}$.

Training process. Following the DDPM framework (Ho, Jain, and Abbeel 2020), we define a timestep $t \in \{1, ..., T\}$ and a variance schedule $\{\beta_1, ..., \beta_T\}$ for diffusion process. Specifically, during the training stage, we corrupt the ground-truth marker latent $\mathbf{z}_{m,0}$ by adding noise at timestep t:

$$\mathbf{z}_{m,t} = \sqrt{\bar{\alpha}_t}\mathbf{z}_{m,0} + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad (1)$$

where $\epsilon \sim \mathcal{N}(0, I)$ is Gaussian noise and $\bar{\alpha}_t = \prod_{k=1}^{t}(1-\beta_k)$. Note that $\mathbf{z}_{m,0}$ is only available during training from ground-truth marker images. This forward process gradually adds noise to the latent representation, allowing the U-Net to learn the denoising operation.

By freezing the SD VAE encoder $\mathcal{E}$ and decoder $\mathcal{D}$ parameters, we train only the conditional diffusion U-Net, which

serves as the denoising network $\hat{\mathbf{v}}_{m,t}(\cdot)$. Specifically, we concatenate a latent pair $[\mathbf{x}, \mathbf{z}_{m,t}]$ along the channel dimension at a timestep t as the input. To accommodate this input, we double the U-Net's input channel size by duplicating weights and halving their values (Ke et al. 2024). We train only the U-Net denoising network $\hat{\mathbf{v}}_{m,t} = \hat{\mathbf{v}}_\theta([\mathbf{x}, \mathbf{z}_{m,t}], t)$ using the following loss function:

$$\mathcal{L}_m = ||\mathbf{v}^*_{m,t} - \hat{\mathbf{v}}_{m,t}||_2^2, \tag{2}$$

where $\mathbf{v}^*_{m,t} = \sqrt{\bar{\alpha}_t}\epsilon - \sqrt{(1-\bar{\alpha}_t)}\mathbf{z}_{m,0}$ represents the training target for noise prediction. This formulation ensures the model learns to accurately predict the noise added during the forward diffusion process.

Inference procedure. During inference, the generation process starts from random noise $\mathbf{z}_{m,T} \sim \mathcal{N}(0, I)$ and iteratively denoises using the learned U-Net. Starting from $t = T$ and stepping down to $t = 1$, we sample:

$$\mathbf{z}_{m,t-1} = \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{z}_{m,t} - \sqrt{1-\bar{\alpha}_t}\hat{\mathbf{v}}_{m,t}([\mathbf{x}, \mathbf{z}_{m,t}], t, c_m). \tag{3}$$

Note that we do not use ground-truth $\mathbf{z}_{m,0}$ during inference. After reaching $t = 0$, the final denoised latent $\hat{\mathbf{z}}_{m,0}$ is reconstructed into pixel space via the frozen VAE decoder: $\hat{\mathbf{I}}_m = \mathcal{D}(\hat{\mathbf{z}}_{m,0})$. The reconstructed image is averaged across channels to produce a single-channel marker image prediction.

Marker-wise conditioning. To enable multiplex staining generation, we condition the diffusion U-Net with marker-specific embedding vectors as depicted in Fig. 2(a). This approach avoids the computational inefficiency of training separate models for each marker type. Moreover, as depicted in Fig. 2(b), we replicate the H&E image latent $\mathbf{x}$ to the number of marker types M, allowing for balanced parameter update across all marker types during training. Correspondingly, we modify the loss function $\mathcal{L}_m$ as

$$\mathcal{L}_M = \frac{1}{M}\sum_{m=1}^{M}\mathcal{L}_m = \frac{1}{M}\sum_{m=1}^{M}||\mathbf{v}^*_{m,t} - \hat{\mathbf{v}}_{m,t}||_2^2, \tag{4}$$

where $\mathbf{v}^*_{m,t}$ and $\hat{\mathbf{v}}_{m,t} = \hat{\mathbf{v}}_\theta([\mathbf{x}, \mathbf{z}_{m,t}], t, c_m)$ denote the optimization target and prediction for marker type m, respectively, with c_m denoting marker-wise conditioning.

Inspired by Fu et al. (2024), we implement marker-wise conditioning by representing each marker type as a one-hot vector. Unlike text conditioning, which can be ambiguous with many marker types, we adopt marker-wise one-hot embedding to provide distinct and scalable signals for multi-marker generation. The one-hot vector undergoes positional encoding and is then element-wise added to the time embedding to condition the model on the specific marker type m. This technique enables the diffusion U-Net to efficiently learn marker-specific features, allowing for multiple marker image generation. A detailed analysis is provided in Sec. 4.

3.2 Fine-tuning for Color Contrast Fidelity

While our model achieves multiplexed image generation in Sec. 3.1, color contrast fidelity remains a primary challenge due to dataset bias toward dark background regions. The visual comparison is detailed in Sec. 4. Moreover, efficient processing can also be important for large-scale pathology data, as diffusion models are intrinsically slow due to their iterative process. While approaches such as DDIM (Song, Meng, and Ermon 2020) help accelerate inference, other techniques like test-time ensembling (Ke et al. 2024) introduce computational overhead for improved performance.

To address both color fidelity and inference efficiency, we introduce a second, fine-tuning stage as shown in Fig. 2(c). In this stage, inspired by recent latent diffusion advances (Garcia et al. 2024), we enable single-step inference by fixing $t = T$ and replacing the random noise with zero noise ($\epsilon = 0$) , optimizing the model directly in pixel space. This modification allows for rapid, deterministic image generation and, crucially, supports the direct application of task-specific pixel-level supervision for improved color contrast. Consistent with previous training phases, we update only the parameters of the diffusion U-Net, while keeping the VAE encoder and decoder fixed.

We employ a pixel-level fine-tuning loss $\mathcal{L}_{FT}$ as follows:

$$\mathcal{L}_{FT} = \frac{1}{M}\sum_{m=1}^{M}[(1-\lambda)\|\mathbf{I}^*_m - \hat{\mathbf{I}}_m\|_1 + \lambda\|\mathbf{I}^*_m - \hat{\mathbf{I}}_m\|_2^2], \tag{5}$$

where $\hat{\mathbf{I}}_m$ and $\mathbf{I}^*_m$ denote the prediction and ground truth for marker type m, respectively. And we combine the L_1 loss $\|\cdot\|_1$ and L_2 loss $\|\cdot\|_2^2$ at the pixel level after passing through the fixed decoder $\mathcal{D}$. By directly applying this combined loss at the pixel level rather than the latent space, we significantly enhance the marker-specific color fidelity and overall visual quality. Details and further analysis of the hyperparameter λ within Eq. (5) can be found in Sec. 4.

4 Experiments

4.1 Setup

Datasets. We utilized two public datasets, HEMIT (Bian et al. 2024) and Orion-CRC (Lin et al. 2023), each providing paired H&E and multiplex-stained images, thus offering a robust benchmark for evaluating virtual multiplex staining techniques. **(1) HEMIT** contains H&E images and three mIHC markers: DAPI (highlighting cell nuclei), pan-cytokeratin (panCK, identifying tumor regions), and CD3 (marking T cells), all crucial for analyzing the tumor microenvironment and immune activity. The dataset consists of 5,292 1024×1024 pixel image pairs with 50% overlap, split into training (3,717 pairs), validation (630 pairs), and test (945 pairs) sets. For rigorous evaluation, we curated 292 image pairs from the test set by excluding spatially overlapping regions and filtering out empty patches that can hinder meaningful quantitative evaluation. This strategy prevents artificial inflation of metrics (*e.g.*, PSNR values above 60 dB) observed in biologically irrelevant, low-signal regions.
(2) Orion-CRC comprises 41 colon cancer whole slide images (WSIs), each with 18 IF marker channels paired with H&E stained sections. These include Hoechst (for nuclei), as well as markers capturing immune, epithelial, and cell state features, all informative for spatial tumor characterization.

Method	SSIM				R				PSNR (dB)			
	DAPI	CD3	panCK	Avg.	DAPI	CD3	panCK	Avg.	DAPI	CD3	panCK	Avg.
pix2pix (Isola et al. 2017)	0.750	0.879	0.572	0.734	0.901	0.381	0.586	0.623	31.53	24.82	26.29	27.55
pix2pixHD (Wang et al. 2018)	0.702	0.844	0.579	0.709	0.948	0.592	0.726	0.755	33.36	26.14	28.07	29.19
HEMIT (Bian et al. 2024)	0.769	0.882	0.659	0.770	0.964	0.567	0.707	0.746	32.82	25.90	27.63	28.78
Parmar et al. (2024)	0.606	0.845	0.652	0.701	0.924	0.332	0.595	0.617	32.15	24.09	26.07	27.44
Marigold (Ke et al. 2024)	0.801	0.711	0.546	0.686	0.956	0.581	0.711	0.750	33.91	26.04	28.13	29.36
Ours	**0.855**	**0.889**	**0.763**	**0.836**	**0.972**	**0.633**	**0.781**	**0.795**	**35.40**	**26.53**	**29.86**	**30.60**

Table 1: Quantitative comparison on the HEMIT dataset (**Bold**: Best, Underline: Second best).

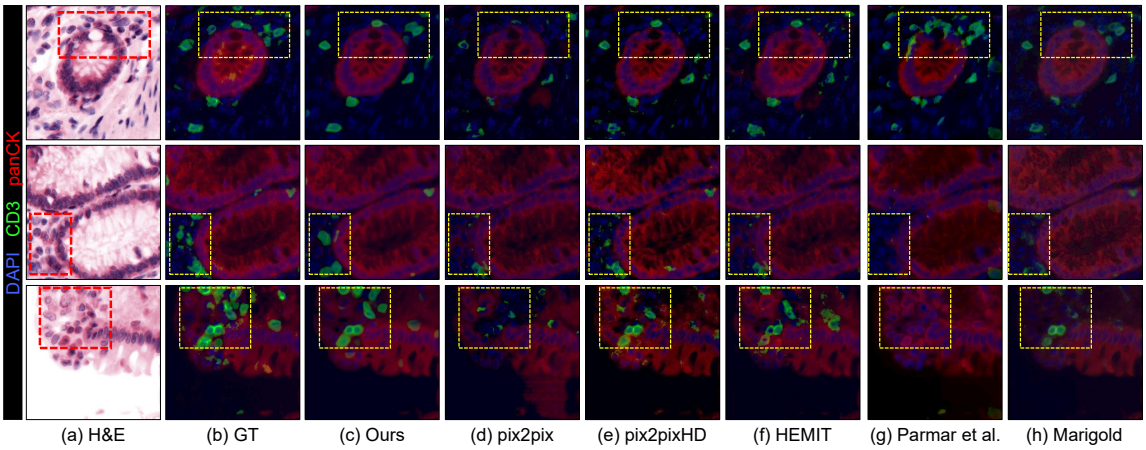

Figure 3: Qualitative comparison on the HEMIT dataset. Each marker type is visualized in distinct colors.

WSIs (average 81.6G pixels) were cropped into 512×512 patches. After filtering out low-tissue-content patches, the dataset consists of 132,610 training patches (31 WSIs) and 39,625 test patches (10 WSIs). For quantitative analysis, patches with an SSIM > 0.8 to an empty array were excluded for each marker to avoid metric inflation due to biologically irrelevant regions. Due to marker rarity, fewer than 1,000 patches were available for CD31, FOXP3, and CD8a.

Evaluation metrics. We use the Structural Similarity Index Measure (SSIM), Pearson correlation score (R), and PSNR, following previous works (Burlingame et al. 2020; Bian et al. 2024).

Implementation details. We use SD v2 as the backbone (Rombach et al. 2022), with patch sizes of 512×512. Our models were trained and fine-tuned on four NVIDIA H100 GPUs following the settings of Ke et al. (2024) and Garcia et al. (2024). Further details are provided in the supplementary material.

Baselines. We compare our approach to several state-of-the-art methods, including pix2pix (Isola et al. 2017), pix2pixHD (Wang et al. 2018), HEMIT (Bian et al. 2024), Parmar et al. (2024), and Marigold (Ke et al. 2024). For the HEMIT dataset with 3 output channels, we implemented Marigold by injecting 3-channel mIHC images into the SD VAE encoder. pix2pixHD and Marigold were excluded from Orion-CRC evaluations due to architectural limitations for multi-channel (18-marker) generation. Pix2pix and HEMIT were adapted to output 18 marker channels for the Orion-CRC dataset. We retrained HEMIT on its own dataset after observing performance discrepancies with the authors' released checkpoint. Parmar et al. was excluded from Orion-CRC evaluations due to severe mode collapse.

4.2 Results

Performance comparison. In Table 1, on the HEMIT dataset, our approach achieves the highest average SSIM (+0.066), R (+0.040), and PSNR (+1.238) compared to the second-highest scores, achieving the highest in every marker and metric. Figure 3 further highlights the advantages of our approach, particularly in capturing CD3 marker signals; for visualization, each marker type is shown in a distinct color. (h) Marigold often fails to localize panCK signals (row 3) and CD3 signals (yellow-box regions), while (g) Parmar et al. tend to over-predict CD3 signals (row 1) or fail to detect CD3 signals (row 3). (f) HEMIT shows better results than other methods, but our method consistently provides clearer and more accurate localizations, aligning with the quantitative results in Table 1.

On Orion-CRC, Table 2 demonstrates that our approach achieved the highest average scores, with improvements of +0.039 SSIM, +0.117 R, and +1.358 PSNR compared to the second-best results, achieving the highest SSIM for 13, R for 15, and PSNR for 18 out of 18 markers. For markers with sufficient signal presence, such as Hoechst (cellular structures), all methods performed well. HEMIT achieved the second-highest performance on Hoechst, but generally un-

Method	Metric	Hoechst	AF1	CD31	CD45	CD68	CD4	FOXP3	CD8a	CD45RO	CD20	PD-L1	CD3e	CD163	E-cad.	PD-1	Ki67	panCK	SMA	Avg.
pix2pix	SSIM	0.751	**0.866**	0.776	0.709	0.779	0.709	**0.638**	0.654	0.739	**0.755**	0.692	0.696	0.713	0.680	0.754	0.710	0.732	0.673	0.724
	R	0.891	0.473	**0.096**	0.427	0.173	0.097	**0.068**	0.096	0.304	0.180	0.205	0.308	0.088	0.434	0.103	0.309	0.531	0.213	0.277
	PSNR	26.54	42.27	37.83	31.34	36.91	39.13	32.05	30.45	36.61	37.92	33.00	30.87	30.72	31.15	36.21	32.86	29.67	31.12	33.70
HEMIT	SSIM	0.762	0.835	0.762	0.590	0.726	0.654	0.622	**0.715**	0.752	0.578	0.726	0.682	**0.721**	0.629	0.700	0.660	0.613	0.683	0.690
	R	0.902	**0.533**	-0.001	0.256	0.000	-0.001	0.000	0.002	0.343	0.000	0.050	0.143	0.000	0.281	0.071	0.262	0.221	0.000	0.170
	PSNR	26.86	42.77	37.85	30.06	36.56	38.69	32.09	28.70	37.54	36.74	32.05	31.47	31.06	30.56	36.15	32.44	28.33	34.02	33.55
Ours	SSIM	**0.779**	0.852	**0.785**	**0.770**	**0.815**	**0.803**	0.623	0.712	**0.838**	0.718	**0.753**	**0.724**	0.718	**0.771**	**0.764**	**0.793**	**0.747**	**0.761**	**0.763**
	R	**0.912**	0.510	0.027	**0.563**	**0.319**	**0.340**	-0.053	**0.232**	**0.480**	**0.290**	**0.359**	**0.421**	**0.264**	**0.669**	**0.127**	**0.608**	**0.637**	**0.394**	**0.394**
	PSNR	**27.77**	**43.30**	**38.19**	**32.94**	**37.38**	**40.62**	**32.15**	**32.24**	**38.65**	**38.01**	**34.62**	**32.95**	**31.82**	**32.97**	**36.76**	**34.39**	**30.61**	**35.72**	**35.06**

Table 2: Quantitative comparison on the Orion-CRC dataset (**Bold**: best, Underline: second best).

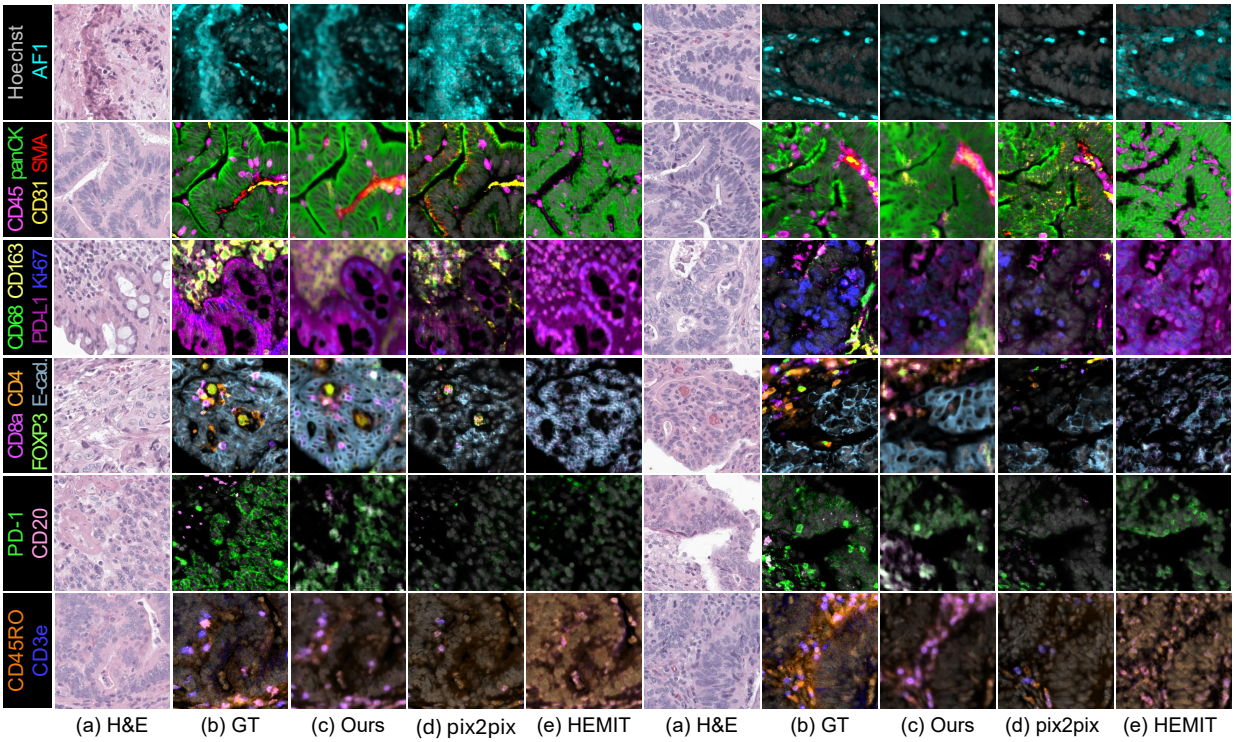

Figure 4: Qualitative comparison on Orion-CRC. Each row depicts different IF markers in various colors. Columns show (a) H&E, (b) ground truth IF (GT), and (c-e) virtual IF by Ours, pix2pix, and HEMIT, respectively.

derperformed compared to pix2pix on other marker types, with pix2pix consistently showing the second-best performance overall. Our method showed lower R scores for CD31 and FOXP3, likely a consequence of the limited number of available patches for these markers (<1,000). Nevertheless, as shown in Fig. 4, our model does not completely miss these signals and still provides reasonable localization (CD31 in row 2, FOXP3 in row 4); in contrast, other methods frequently miss signals for certain marker types such as SMA (row 2) and CD20 (row 5). While our results exhibit some mild blurriness and tend to predict certain markers over broader regions compared to the ground truth (*e.g.*, panCK in row 2, CD68 and CD163 in row 3), this tendency is likely due to structural properties of the latent diffusion model, in particular the absence of skip connections and its operation in the latent space, as previously observed in (Parmar et al. 2024). For the remaining marker types, our method generally matches or exceeds the performance of previous approaches, yielding consistent improvements in both localization accuracy and overall metric scores.

Overall, our method achieves superior localization and performance compared to previous methods, highlighting its potential for H&E staining analysis supported by virtual multiplex staining.

Ablation studies. Table 3 highlights the scalability and effectiveness of marker-wise one-hot embedding approach

Dataset		Condition	SSIM	R	PSNR (dB)
HEMIT		Text	0.667	**0.772**	30.09
	✓	One-hot	**0.673**	0.770	**30.21**
Orion-CRC		Text	0.288	-0.003	18.01
	✓	One-hot	**0.662**	**0.371**	**30.66**

Table 3: Ablation on marker type conditioning strategy

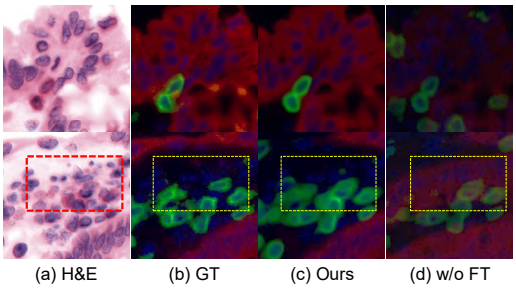

Figure 5: Color fidelity comparison on the HEMIT dataset: (a) H&E, (b) ground truth, (c) Ours, (d) without fine-tuning.

Method	Run time (sec/sample)	Memory (MB)	SSIM	R	PSNR (dB)
(DDIM steps: 50, Ensemble: 10 (Ke et al. 2024))					
Single-plex	117.18	14607	0.670	0.755	30.32
Multiplex	110.59	33838	0.673	0.770	30.21
(DDIM steps: 1, Ensemble: 1)					
Multiplex	0.13	7850	0.757	0.760	29.38
+Fine-tune	0.13	7850	**0.836**	**0.795**	**30.60**

Table 4: Ablation on the performance and inference cost on the HEMIT dataset.

compared to text-based conditioning, with these results obtained prior to the fine-tuning stage. On the HEMIT dataset (3 marker types), both strategies produced comparable results, with one-hot conditioning yielding a slightly higher SSIM and PSNR. However, on Orion-CRC (18 marker types), one-hot conditioning achieved a substantial gain (SSIM 0.662, PSNR 30.663), whereas text conditioning failed to generate outputs aligned with the specified marker types, resulting in much lower SSIM (0.288) and PSNR (18.014). This confirms that while text-based conditioning can work in small-marker scenarios (like HEMIT), it does not scale to practical multiplex settings with numerous marker types. In summary, the ablation demonstrates that marker-wise one-hot embedding is a robust and scalable conditioning strategy for virtual multiplex staining.

Figure 5 visually compares our fine-tuned model (c) and the non-finetuned model (d) w/o FT. The model without fine-tuning (d) exhibits significant color distortion and struggles to accurately reproduce the marker-specific colors in the ground truth (b) GT. Additionally, it shows poor localization accuracy, particularly evident in row 2, with false positive panCK (red) predictions in the yellow box. In contrast, our fine-tuned model (c) achieves superior color fidelity that closely matches the ground truth and demonstrates highly accurate spatial localization of marker signals. This comparison emphasizes the importance of incorporating pixel-level loss-based fine-tuning to enhance both color fidelity and overall model performance.

Dataset		λ	SSIM	R	PSNR (dB)
HEMIT		0	**0.837**	0.788	30.71
	✓	0.5	0.836	0.795	30.60
		1.0	0.803	**0.812**	**31.26**
Orion-CRC		0	0.726	0.255	34.39
		0.5	0.741	0.283	34.58
	✓	1.0	**0.763**	**0.394**	**35.06**

Table 5: Ablation on loss weight λ in fine-tuning stage

Table 4 summarizes the impact of the fine-tuning stage on model performance and inference cost on the HEMIT dataset. A single-plex model (one model per marker) serves as baseline but is impractical for virtual multiplex staining with a large number of markers, requiring multiple models. Our multiplex model generates all markers in a single network, improving training efficiency and practical scalability. We used 10-fold test-time ensembling (Ke et al. 2024) for optimal performance. Multiplex model shows slightly higher SSIM and R, but with much heavier memory demand. Notably, single-step inference in the multiplex model yields even higher SSIM (0.757) than 50-step sampling with ensembling (0.673). After the fine-tuning stage, our approach achieves the highest performance with dramatically reduced runtime and memory, demonstrating that fine-tuning is crucial for both image fidelity and efficient, scalable virtual multiplex staining.

Table 5 summarizes the effect of the fine-tuning loss weight λ. For HEMIT, we selected $\lambda = 0.5$, prioritizing SSIM and R. Notably, our method outperformed all related works in Table 1 for any tested λ. For Orion-CRC, $\lambda = 1.0$ was chosen as it achieved the best overall metric scores. This variation in optimal λ, potentially reflecting differences in marker signal distribution between datasets, suggests that dataset adaptive loss function selection may be beneficial for maximizing performance in multiplex settings.

Discussion. While our framework achieves superior performance for virtual multiplex staining, several challenges remain. Single-step inference reduces per-sample cost, but computational cost scales with marker count, which is an inherent challenge for high-dimensional applications. Our model also lacks explicit inter-marker correlation modeling, potentially limiting image fidelity. Addressing these points is crucial for practical deployment.

5 Conclusion

In this paper, we proposed a novel framework for virtual multiplex staining using marker-wise conditioned diffusion models. By conditioning on marker-specific embeddings and refining with pixel-level loss functions, we achieved state-of-the-art performance on public datasets while improving practical applicability. Future directions include enhancing structural detail, explicitly modeling inter-marker correlations, and evaluating clinical impact in pathology.

Acknowledgements

This work was supported in part by the National Research Foundation of Korea under Grant RS-2024-00349697 and Grant RS-2021-NR060143; in part by the Institute for Information and Communications Technology Planning and Evaluation under Grant IITP-2025-RS-2020-II201819; in part by the Technology Development Program funded by the Ministry of SMEs and Startups (MSS), South Korea, under Grant RS-2024-00437796; in part by the National Research Council of Science and Technology (NST) grant funded by Korean Government [Ministry of Science and Information and Communications Technology (MSIT)] under Grant GTL24031-000; and in part by Korea University Grant.

References

Bian, C.; Philips, B.; Cootes, T.; and Fergie, M. 2024. HEMIT: H&E to Multiplex-immunohistochemistry Image Translation with Dual-Branch Pix2pix Generator. *arXiv preprint arXiv:2403.18501*.

Burlingame, E. A.; McDonnell, M.; Schau, G. F.; Thibault, G.; Lanciault, C.; Morgan, T.; Johnson, B. E.; Corless, C.; Gray, J. W.; and Chang, Y. H. 2020. SHIFT: speedy histological-to-immunofluorescent translation of a tumor signature enabled by deep learning. *Scientific reports*, 10(1): 17507.

Chen, F.; Zhang, R.; Zheng, B.; Sun, Y.; He, J.; and Qin, W. 2024. Pathological semantics-preserving learning for H&E-to-IHC virtual staining. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 384–394. Springer.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Dubey, S.; Chong, Y.; Knudsen, B.; and Elhabian, S. Y. 2024. VIMs: Virtual Immunohistochemistry Multiplex staining via Text-to-Stain Diffusion Trained on Uniplex Stains. In *International Workshop on Machine Learning in Medical Imaging*, 143–155. Springer.

Fu, X.; Yin, W.; Hu, M.; Wang, K.; Ma, Y.; Tan, P.; Shen, S.; Lin, D.; and Long, X. 2024. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *European Conference on Computer Vision*, 241–258. Springer.

Garcia, G. M.; Zeid, K. A.; Schmidt, C.; de Geus, D.; Hermans, A.; and Leibe, B. 2024. Fine-Tuning Image-Conditional Diffusion Models is Easier than You Think. *arXiv preprint arXiv:2409.11355*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

Ke, B.; Obukhov, A.; Huang, S.; Metzger, N.; Daudt, R. C.; and Schindler, K. 2024. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9492–9502.

Latonen, L.; Koivukoski, S.; Khan, U.; and Ruusuvuori, P. 2024. Virtual staining for histology by deep learning. *Trends in Biotechnology*.

Li, F.; Hu, Z.; Chen, W.; and Kak, A. 2023. Adaptive supervised patchnce loss for learning h&e-to-ihc stain translation with inconsistent groundtruth image pairs. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 632–641. Springer.

Lin, J.-R.; Chen, Y.-A.; Campton, D.; Cooper, J.; Coy, S.; Yapp, C.; Tefft, J. B.; McCarty, E.; Ligon, K. L.; Rodig, S. J.; et al. 2023. High-plex immunofluorescence imaging and traditional histology of the same tissue section for discovering image-based biomarkers. *Nature cancer*, 4(7): 1036–1052.

Parmar, G.; Park, T.; Narasimhan, S.; and Zhu, J.-Y. 2024. One-step image translation with text-to-image models. *arXiv preprint arXiv:2403.12036*.

Pati, P.; Karkampouna, S.; Bonollo, F.; Compérat, E.; Radić, M.; Spahn, M.; Martinelli, A.; Wartenberg, M.; Kruithof-de Julio, M.; and Rapsomaniki, M. 2024. Accelerating histopathology workflows with generative AI-based virtually multiplexed tumour profiling. *Nature Machine Intelligence*, 1–17.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2256–2265. pmlr.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Sood, A.; Miller, A. M.; Brogi, E.; Sui, Y.; Armenia, J.; McDonough, E.; Santamaria-Pang, A.; Carlin, S.; Stamper, A.; Campos, C.; et al. 2016. Multiplexed immunofluorescence delineates proteomic cancer cell states associated with metabolism. *JCI insight*, 1(6).

Tan, W. C. C.; Nerurkar, S. N.; Cai, H. Y.; Ng, H. H. M.; Wu, D.; Wee, Y. T. F.; Lim, J. C. T.; Yeong, J.; and Lim, T. K. H. 2020. Overview of multiplex immunohistochemistry/immunofluorescence techniques in the era of cancer immunotherapy. *Cancer Communications*, 40(4): 135–153.

Tumanyan, N.; Geyer, M.; Bagon, S.; and Dekel, T. 2023. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1921–1930.

Wang, T.; Wang, M.; Wang, Z.; Wang, H.; Xu, Q.; Cong, F.; and Xu, H. 2025. ODA-GAN: Orthogonal Decoupling Alignment GAN Assisted by Weakly-supervised Learning for Virtual Immunohistochemistry Staining. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 25920–25929.

Wang, T.-C.; Liu, M.-Y.; Zhu, J.-Y.; Tao, A.; Kautz, J.; and Catanzaro, B. 2018. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8798–8807.

Wharton Jr, K. A.; Wood, D.; Manesse, M.; Maclean, K. H.; Leiss, F.; and Zuraw, A. 2021. Tissue multiplex analyte detection in anatomic pathology–pathways to clinical implementation. *Frontiers in molecular biosciences*, 8: 672531.

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

Zhao, W.; Rao, Y.; Liu, Z.; Liu, B.; Zhou, J.; and Lu, J. 2023. Unleashing text-to-image diffusion models for visual perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5729–5739.