PALETTE: Pathology-Aware Latent Embedding for Targeted Tissue Expression

Rongze Ma[1,2⋆], Mengkang Lu[1,2⋆], Yihang Fu[1,2], Xinke Ma[1,2], and Yong Xia[1,2(✉)]

[1] Ningbo Institute of Northwestern Polytechnical University, Ningbo 315048, China
[2] National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China
yxia@nwpu.edu.cn

Abstract. Virtual staining aims to computationally synthesize immunohistochemistry (IHC) images from routine Hematoxylin and Eosin (H&E) slides, offering a scalable and cost-effective alternative to chemical staining. However, accurate virtual IHC generation remains challenging due to the complex and partially decoupled relationship between morphology and molecular expression, as well as spatial misalignment in paired training data. Most existing approaches rely on reconstruction-oriented encoder–decoder architectures trained from scratch, compressing high-resolution pathology images into low-dimensional latent spaces that may discard semantically critical structural information. In this work, we reformulate virtual staining as a representation-decoupled cross-modal generation problem and propose PALETTE, a foundation-guided framework that separates semantic encoding from molecular synthesis. A frozen pathology foundation model (PathFM) encodes H&E images into high-capacity continuous representations, which serve as multi-scale conditions for a Cond-VAR that progressively generates discrete IHC tokens in a coarse-to-fine manner. By decoupling semantic representation from texture reconstruction, PALETTE mitigates information bottlenecks and reduces sensitivity to latent misalignment, promoting structurally coherent molecular synthesis. Experiments on IHC4BC demonstrate state-of-the-art structural and perceptual fidelity across multiple biomarkers, while downstream HER2 classification and ER scoring confirm preservation of diagnostically relevant patterns.

Keywords: Virtual Staining · Visual Autoregressive Transformer.

1 Introduction

Immunohistochemistry (IHC) is central to modern diagnostic pathology, enabling visualization of protein expression at cellular resolution and supporting tumor subtyping, prognostic stratification, and therapeutic decision-making [15,21].

⋆ R. Ma and M. Lu contributed equally. Corresponding author: Y. Xia.

In contrast, routine Hematoxylin and Eosin (H&E) staining primarily captures tissue morphology without explicit molecular specificity. Although indispensable in clinical workflows, chemical IHC staining is labor-intensive, time-consuming, and limited by tissue availability. Virtual staining, which computationally synthesizes IHC images from H&E slides, has therefore emerged as a promising strategy to reduce cost and improve workflow efficiency [3,18]. Despite recent advances, high-fidelity virtual IHC generation remains challenging. The mapping from morphology to molecular expression is complex, context-dependent, and only partially observable from H&E appearance alone. Accurate biomarker localization requires integrating fine-grained cellular morphology with global tissue organization. Moreover, paired H&E–IHC datasets frequently exhibit spatial misalignment due to sectioning variability and imperfect registration, rendering strict pixel-level supervision unreliable [20].

Early approaches treated virtual staining as a generic image-to-image translation task, employing GAN- and diffusion-based architectures such as Pix2Pix [16], CycleGAN [33], CUT [24], HistDist [14], and SynDiff [23]. While effective for natural images, these models largely frame H&E-to-IHC translation as low-level appearance transformation without explicitly modeling biological dependencies. Subsequent pathology-specific methods introduce patch-level consistency constraints [17,31] or protein-aware objectives [5] to mitigate misalignment. Variational [30,12] and autoregressive formulations [26] further attempt cross-modal synthesis via shared latent representations. However, despite architectural diversity, most existing methods share a common limitation: they rely on task-specific encoders trained from scratch to compress high-resolution histological structures into low-capacity latent spaces optimized for reconstruction or local alignment [32]. Such compression risks discarding long-range tissue context and semantically meaningful structural cues essential for coherent molecular inference [9,6]. Consequently, these models remain sensitive to spatial distortions and often exhibit structural inconsistencies in complex or misaligned regions [4,10]. The fundamental issue lies not merely in the generative backbone, but in the representation paradigm itself.

Large-scale pathology foundation models, such as CONCH [19] and UNI [8], offer an alternative. Pretrained on extensive whole-slide image collections, these models encode semantically structured and spatially coherent histomorphological representations that preserve multi-scale context [7]. Leveraging such pretrained encoders as frozen representation backbones enables a different strategy for virtual staining: generation can be conditioned on high-capacity, semantically faithful features rather than learned compression bottlenecks. This representation–generation decoupling provides a principled basis for spatially consistent and biologically plausible synthesis.

In this work, we revisit virtual staining from a representation-decoupling perspective and propose PALETTE, a foundation-guided cross-modal generation framework that separates semantic encoding from molecular texture synthesis. A frozen pathology foundation model extracts continuous, high-capacity representations from H&E images, which serve as multi-scale structural conditions for a

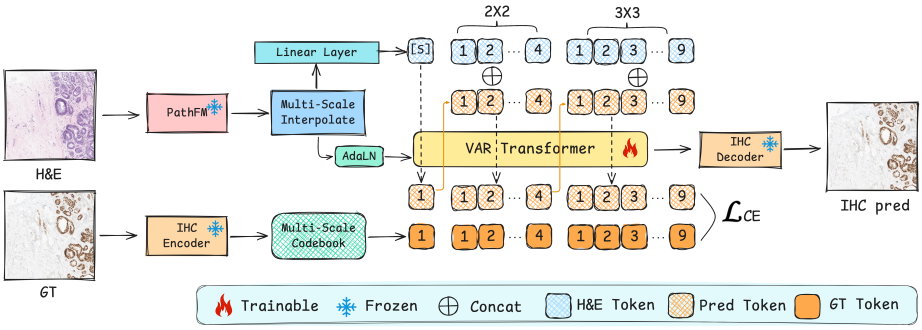

Fig. 1. The overall architecture of the proposed PALETTE framework. A frozen foundation model first extracts multi-scale continuous features (V_{HE}) from the source H&E image as robust structural anchors. Concurrently, a modality-specific VQGAN compresses the target IHC image into a hierarchical discrete token space (Z). Finally, a Cond-VAR bridges the two modalities by autoregressively predicting the discrete IHC tokens in a coarse-to-fine sequence, guided by the V_{HE} condition via a dual-pathway strategy (AdaLN and spatial concatenation) before VQGAN reconstruction.

conditional Visual Autoregressive Transformer (Cond-VAR). The VAR progressively predicts discrete IHC tokens in a coarse-to-fine manner, preserving source semantic structure while modeling molecular appearance autoregressively.

Our contributions are threefold: (1) We formulate virtual IHC generation as a representation-decoupled cross-modal synthesis problem and introduce a framework that leverages a frozen PathFM for semantically rich H&E encoding. (2) We design a Cond-VAR guided by multi-scale continuous conditions, enabling structured molecular token prediction without heavily compressed latent alignment. (3) Through evaluation across multiple biomarkers and downstream HER2 classification and ER clinical scoring tasks, we demonstrate improved structural fidelity and preservation of diagnostically relevant patterns.

2 Method

The proposed PALETTE framework decouples cross-modal virtual staining into two stages: robust semantic encoding and high-fidelity molecular synthesis. A frozen PathFM first extracts continuous semantic features from H&E images, while a modality-specific vector-quantized autoencoder projects target IHC images into a hierarchical discrete token space. A Cond-VAR then bridges these two spaces by autoregressively predicting discrete IHC tokens conditioned on multi-scale continuous H&E representations.

2.1 Continuous Semantic Representation of the Source Modality

Effective cross-modal generation requires a source representation that faithfully captures complex histomorphological structures across multiple spatial scales.

Conventional compression-based autoencoders rely on aggressive spatial downsampling and reconstruction objectives, producing low-capacity latent spaces obscuring fine-grained nuclear morphology and long-range tissue organization.

To preserve structural and semantic information, we employ a frozen, pretrained PathFM (e.g., CONCH [19]) as the H&E encoder, denoted as $\mathcal{E}_{\text{HE}}$. Rather than learning a task-specific bottleneck representation, this design leverages large-scale pretraining to provide high-capacity, semantically structured features while avoiding additional information loss during cross-modal mapping.

Given an input H&E image $X_{\text{HE}} \in \mathbb{R}^{H\times W\times 3}$, we extract its continuous latent representation:

$$f_{\text{HE}} = \mathcal{E}_{\text{HE}}(X_{\text{HE}}) \in \mathbb{R}^{C\times h\times w}, \tag{1}$$

where C denotes the hidden dimension of the encoder. Unlike the target IHC modality, which is discretized into tokens, the H&E representation remains continuous to retain maximal semantic guidance.

To align the continuous H&E representation with the multi-scale autoregressive decoding process [25], we construct a feature pyramid over predefined spatial scales $S = (s_1, s_2, \ldots, s_K)$ (e.g., $1, 2, 3, 4, 5, 6, 8, 10, 13, 16$). The base feature map f_{HE} is progressively resized using area interpolation:

$$f_{\text{HE}}^{(k)} = \text{Interpolate}(f_{\text{HE}}, s_k, s_k). \tag{2}$$

Each scale-specific feature map $f_{\text{HE}}^{(k)} \in \mathbb{R}^{C\times s_k\times s_k}$ is flattened into a sequence of continuous tokens $v_k \in \mathbb{R}^{s_k^2\times C}$. Concatenating all scales yields a multi-scale conditional sequence:

$$V_{\text{HE}} = [v_1, v_2, \ldots, v_K] \in \mathbb{R}^{L\times C}, \tag{3}$$

where $L = \sum_{k=1}^{K} s_k^2$. During generation, V_{HE} serves as a structural prior. After linear projection, it modulates the VAR Transformer blocks via Adaptive Layer Normalization (AdaLN), guiding scale-by-scale autoregressive prediction of discrete IHC tokens.

2.2 Discrete Representation of Target Modality

To facilitate cross-modal generation via Cond-VAR, we train a modality-specific VQGAN (VQ_{IHC}) [13] to project target IHC images into a discrete token space.

Given an IHC image X_{IHC}, the encoder $\mathcal{E}_{\text{IHC}}$ extracts a high-resolution continuous latent $f \in \mathbb{R}^{h\times w\times d}$. To construct the multi-scale token sequence $Z = (z_1, \ldots, z_K)$ required by VAR, f is progressively down-sampled into K spatial scales. At each scale k, the feature $f_k = \text{Downsample}(f, h_k, w_k)$ is discretized using a shared codebook $\mathcal{Z} \in \mathbb{R}^{N\times d}$ via nearest-neighbor quantization:

$$z_k = \arg\min_{z\in\mathcal{Z}} \|f_k^{(i,j)} - z\|_2,$$

where (i, j) denotes the spatial location and d represents the channel dimension of the features.

To capture fine-grained IHC morphological structures, the decoder $\mathcal{D}_{\text{IHC}}$ reconstructs the image strictly from the highest-resolution token map: $\hat{X}_{\text{IHC}} = \mathcal{D}_{\text{IHC}}(z_K)$. The VQ_{IHC} model is optimized via:

$$\mathcal{L}_{\text{VQ}} = \mathcal{L}_{\text{rec}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{feat}}\mathcal{L}_{\text{feat}},$$

combining l_1 reconstruction ($\mathcal{L}_{\text{rec}}$), patch-based adversarial ($\mathcal{L}_{\text{adv}}$), and GAN feature matching ($\mathcal{L}_{\text{feat}}$) losses to balance pixel accuracy, textural realism, and structural consistency. We empirically set $\lambda_{\text{adv}} = 0.3$ and $\lambda_{\text{feat}} = 0.5$. Once trained, the components are frozen for the subsequent VAR stage. Specifically, $\mathcal{E}_{\text{IHC}}$ and $\mathcal{Z}$ are employed to extract the ground-truth multi-scale tokens, whereas $\mathcal{D}_{\text{IHC}}$ is strictly responsible for reconstructing the high-fidelity IHC images from the generated tokens.

2.3 Cross-Modal Generation via Cond-VAR

Unlike standard Visual Autoregressive (VAR) models conditioned on global class labels, cross-modal H&E to IHC translation requires strict spatial constraints. We introduce a Cond-VAR framework that leverages multi-scale continuous H&E features $V_{\text{HE}} = (v_1, \ldots, v_K)$ to guide the next-scale prediction of discrete IHC tokens $Z = (z_1, \ldots, z_K)$. The conditional likelihood is formulated as:

$$p(Z \mid V_{\text{HE}}) = \prod_{k=1}^{K} p(z_k \mid z_{<k}, v_{\leq k}).$$

Dual-Pathway Conditioning & Masking. To integrate these multi-scale conditions, we employ a dual-pathway strategy. The lowest-resolution H&E token v_1 captures global tissue context and modulates the VAR Transformer blocks via Adaptive Layer Normalization (AdaLN). For finer scales ($k \geq 2$), projected spatial condition tokens v_k are concatenated with the embedded sequence of previously generated IHC tokens, providing strong pixel-wise structural supervision. To enforce a strictly autoregressive generation process and prevent information leakage, we design a scale-causal attention mask for the PALETTE framework. Specifically, when predicting IHC tokens at the current scale k, the mask restricts the attention mechanism such that queries can solely attend to the valid prefix context: the previously generated target IHC tokens from coarser scales ($z_{<k}$) and the corresponding H&E condition tokens up to the current scale ($v_{\leq k}$). This formulation mathematically guarantees that the model is completely blind to any unpredicted target information or conditional structures at finer future scales ($> k$), thereby maintaining a stable, coarse-to-fine cross-modality alignment.

Training and Inference. The model is optimized via cross-entropy loss over all K scales against ground-truth tokens extracted by the frozen $\mathcal{E}_{\text{IHC}}$:

$$\mathcal{L}_{\text{CE}} = -\sum_{k=1}^{K} \log p(z_k \mid z_{<k}, v_{\leq k}).$$

During inference, generation proceeds coarse-to-fine. Initialized by a start token v_1, the Conditional VAR Transformer iteratively predicts next-scale IHC tokens, concatenating them with corresponding spatial conditions v_k until the full-resolution token map z_K is complete. Finally, the frozen decoder yields the IHC image $\hat{X}_{\text{IHC}} = \mathcal{D}_{\text{IHC}}(z_K)$.

3 Experiments and Results

3.1 Datasets and Preprocessing

We evaluate our framework on the public breast cancer IHC4BC [2][3] dataset, which encompasses four clinical biomarkers: Estrogen Receptor (ER) with 59 patients (26,134 patch pairs), Progesterone Receptor (PR) with 60 patients (24,971 pairs), HER2 with 52 patients (20,729 pairs), and Ki67 with 43 patients (20,668 pairs). To ensure reliable pixel-level correspondence, all whole-slide image pairs were registered using DeeperHistReg [29] prior to patch extraction. To standardize the inputs for our network, all extracted patch pairs were resized to 256 × 256 pixels. Furthermore, to rigorously prevent data leakage during evaluation, we partitioned the dataset into fixed training and testing sets at the patient level. For the experiments presented in this work, we utilized a single representative split, allocating approximately 75% of the patients for training and the remaining 25% for testing, strictly balanced by clinical diagnostic labels.

3.2 Implementation Details

All experiments were implemented in PyTorch on RTX 4090 GPUs at a 256 × 256 resolution. The VQ_{IHC} tokenizer (32 base channels, 32-dim latent, 4096 codebook) and a PatchGAN discriminator were trained for 80 epochs ($lr = 2 \times 10^{-4}$). Subsequently, conditions extracted by the frozen PathFM (e.g., CONCH) were linearly projected to 32 dimensions to guide a 16-layer VAR Transformer in autoregressively predicting frozen VQ_{IHC} tokens over 120 epochs.

3.3 Results

Quantitative Evaluation on IHC4BC Dataset We compared our method against general-purpose (Pix2Pix [16], CycleGAN [33], CUT [24], SynDiff [23], MMHVAE [11]) and pathology-specific baselines (ASP [17], PPT [31], PSP-Stain [5]) under identical settings, utilizing the exact same registered image pairs.

As demonstrated in Table 1, PALETTE consistently outperforms state-of-the-art methods across most benchmarks. It achieves the highest SSIM and PSNR on the ER, HER2, and PR subsets, significantly surpassing advanced baselines such as SynDiff and MMHVAE. For the challenging Ki67 subset, PALETTE

[3] https://www.kaggle.com/datasets/akbarnejad1991/ihc4bc-compressed

Table 1. Quantitative comparison on IHC4BC dataset. **Bold** indicates best, underline indicates second-best.

	IHC4BC-ER		IHC4BC-HER2		IHC4BC-PR		IHC4BC-Ki67	
Method	**SSIM ↑**	**PSNR ↑**	**SSIM ↑**	**PSNR ↑**	**SSIM ↑**	**PSNR ↑**	**SSIM ↑**	**PSNR ↑**
Pix2Pix [16]	0.6856	27.4834	0.2690	15.1791	0.6871	27.7055	0.6614	27.8138
CycleGAN [33]	0.6970	25.3865	0.3078	14.1191	0.7030	28.3201	0.6678	27.9537
CUT [24]	0.6771	25.0901	0.3045	14.3601	0.7079	28.9840	0.6243	26.3754
ASP [17]	0.6981	26.4144	0.3102	14.8489	0.7328	29.7684	0.6555	27.7210
PPT [31]	0.6639	26.1911	0.2877	14.8747	0.7217	29.4877	0.6550	27.7547
PSPStain [5]	0.6854	24.8626	0.3069	14.7975	0.6995	28.0571	0.6549	27.1000
SynDiff [23]	0.7375	28.7356	0.3117	14.6236	0.7777	31.5956	0.6827	28.6649
MMHVAE [11]	0.7096	27.1730	0.3057	14.7450	0.7922	31.5219	**0.6905**	28.3483
PALETTE	**0.7589**	**29.5958**	**0.3423**	**16.5205**	**0.8142**	**32.4510**	0.6741	**29.9472**

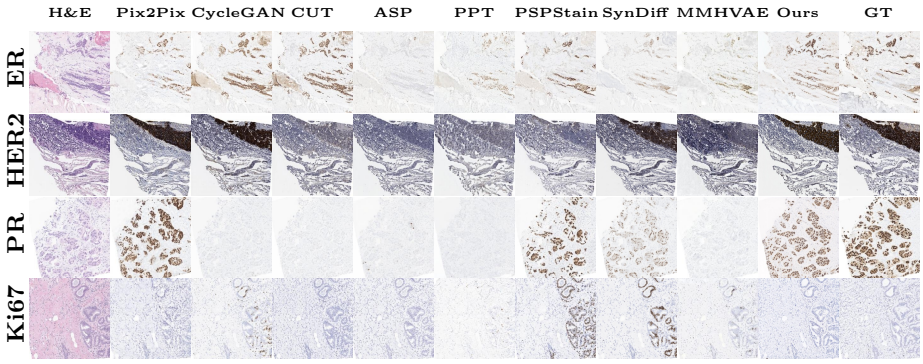

Fig. 2. Visual comparison of ER, HER2, PR and Ki67 generation.

delivers the top PSNR (29.9472), while its SSIM (0.6741) remains highly competitive, trailing SynDiff's 0.6827 by only a narrow margin. These results validate that our foundation-guided framework effectively addresses complex misalignments, ensuring high-fidelity virtual staining (Figure 2).

Clinical Downstream Evaluation We assess the clinical fidelity of PALETTE via HER2 classification [2] and ER scoring tasks [1] (Table 2). PALETTE outperforms all baselines across all metrics, achieving top performance in HER2 classification (ACC: 0.9079, AUC: 0.9626) and ER H-Score regression (R^2: 0.8639, MAE: 15.2620). Unlike competing methods struggling to maintain quantitative biomarkers, our framework faithfully recovers molecular expressions essential for diagnosis, bridging the gap between virtual staining and clinical practice.

3.4 Ablation Studies

We conducted ablation experiments to evaluate the contributions of our architectural designs and the choice of foundation models (Table 3).

Table 2. Quantitative comparison on downstream HER2 classification and ER scoring. **Bold** and underline indicate the best and second-best results, respectively.

	HER2 Classification		ER H-Score		ER Allred-Score	
Method	**ACC ↑**	**AUC ↑**	R^2 ↑	**MAE ↓**	**ACC ↑**	**AUC ↑**
H&E	0.8394	0.8501	0.7355	21.4708	0.4327	0.8500
Pix2Pix [16]	0.8539	0.9214	<u>0.8142</u>	<u>17.5861</u>	<u>0.4476</u>	<u>0.8662</u>
CycleGAN [33]	0.8634	0.9258	0.5646	26.5349	0.3756	0.8073
CUT [24]	0.8363	0.8997	0.4331	29.4779	0.3592	0.7922
ASP [17]	0.8450	0.9129	0.4874	28.0001	0.3789	0.7941
PPT [31]	0.8479	0.9188	0.2853	32.2835	0.3299	0.7412
PSPStain [5]	<u>0.8788</u>	<u>0.9424</u>	0.4860	27.9597	0.3880	0.7974
SynDiff [23]	0.8312	0.9103	0.7010	22.6980	0.4004	0.8314
MMHVAE [11]	0.8435	0.9093	0.4891	27.5740	0.3821	0.8066
PALETTE	**0.9079**	**0.9626**	**0.8639**	**15.2620**	**0.4995**	**0.8857**

Table 3. Ablation study of the proposed PALETTE framework.

Components	**SSIM ↑**	**PSNR ↑**	**Model Variant**	**SSIM ↑**	**PSNR ↑**
w/o AdaLN	0.7163	27.6160	w/ DINOv2 [22]	0.7311	28.6063
w/o VAR	0.6668	13.9647	w/ Virchow [28]	0.7319	28.1128
w/o Condition	0.6571	18.5512	w/ UNI [8]	0.7552	28.9693
w/o PathFM	0.7279	27.1018	w/ CONCH [19]	**0.7589**	**29.5958**

Architectural Components. The ablation of core modules reveals their critical roles in the PALETTE framework. Removing the VAR component results in a drastic performance collapse, with PSNR dropping significantly to 13.9647, confirming that VAR is essential for modeling complex spatial dependencies in virtual staining. Similarly, excluding the condition (w/o Condition) leads to the lowest SSIM (0.6571), highlighting its necessity for structural guidance. The absence of AdaLN (w/o AdaLN) also leads to a noticeable degradation in structural and perceptual metrics, validating its effectiveness in global feature modulation. Furthermore, replacing the frozen PathFM with a trainable VQ encoder [27] (w/o PathFM) leads to inferior results (SSIM: 0.7279), suggesting that standard VQ-based compression may lose crucial histopathological semantic information.
Foundation Model Variants. We further explored the impact of different visual representations. Evaluating various foundation models as encoders shows that domain-specific models (e.g., Virchow [28], UNI [8], CONCH [19]) generally outperform general-purpose models like DINOv2 [22]. Among these, the CONCH-guided variant achieves the best performance (SSIM: 0.7589, PSNR: 29.5958), demonstrating that rich semantic features from specialized foundation models significantly enhance the fidelity of cross-modality stain generation.

4 Discussion

We present PALETTE, a representation-decoupled virtual immunohistochemistry framework. Leveraging a frozen PathFM separates semantic encoding from

texture generation, avoiding compressed latent bottlenecks to achieve high structural fidelity and clinically relevant biomarker preservation. Despite patch-level limitations and autoregressive inference costs, PALETTE establishes a principled representation-driven paradigm for cross-modal pathology synthesis. Future work targets acceleration and whole-slide integration.

Acknowledgments. This work was supported in part by the "Pioneer" and "Leading Goose" R&D Program of Zhejiang, China, under Grant 2025C01201(SD2), in part by the National Natural Science Foundation of China under Grant 92470101, in part by the Project of National Key Clinical Specialty (Department of Medical Imaging, No. 2024017), in part by the Ningbo Key Laboratory of Digital Imaging and Medical-Engineering Interdisciplinarity (No. 2024031).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ahmad Fauzi, M.F., Wan Ahmad, W.S.H.M., Jamaluddin, M.F., Lee, J.T.H., Khor, S.Y., Looi, L.M., Abas, F.S., Aldahoul, N.: Allred scoring of er-ihc stained whole-slide images for hormone receptor status in breast carcinoma. Diagnostics **12**(12), 3093 (2022)
2. Akbarnejad, A., Ray, N., Barnes, P.J., Bigras, G.: Predicting ki67, er, pr, and HER2 statuses from h&e-stained breast cancer images. arXiv preprint arXiv:2308.01982 (2023)
3. Anglade, F., Milner Jr, D.A., Brock, J.E.: Can pathology diagnostic services for cancer be stratified and serve global health? Cancer **126**, 2431–2438 (2020)
4. Bai, B., Yang, X., Li, Y., Zhang, Y., Pillar, N., Ozcan, A.: Deep learning-enabled virtual histological staining of biological samples. Light: Science & Applications **12**(1), 57 (2023)
5. Chen, F., Zhang, R., Zheng, B., Sun, Y., He, J., Qin, W.: Pathological semantics-preserving learning for h&e-to-ihc virtual staining. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 384–394. Springer (2024)
6. Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., Mahmood, F.: Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16144–16155 (2022)
7. Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., Mahmood, F.: Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16144–16155 (2022)
8. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature medicine **30**(3), 850–862 (2024)
9. Cohen, J.P., Luck, M., Honari, S.: Distribution matching losses can hallucinate features in medical image translation. In: International conference on medical image computing and computer-assisted intervention. pp. 529–536. Springer (2018)

10. De Haan, K., Zhang, Y., Zuckerman, J.E., Liu, T., Sisk, A.E., Diaz, M.F., Jen, K.Y., Nobori, A., Liou, S., Zhang, S., et al.: Deep learning-based transformation of h&e stained tissues into special stains. Nature communications **12**(1), 4884 (2021)
11. Dorent, R., Haouchine, N., Golby, A., Frisken, S., Kapur, T., Wells, W.: Unified cross-modal medical image synthesis with hierarchical mixture of product-of-experts. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
12. Dorent, R., Haouchine, N., Kogl, F., Joutard, S., Juvekar, P., Torio, E., Golby, A.J., Ourselin, S., Frisken, S., Vercauteren, T., et al.: Unified brain mr-ultrasound synthesis using multi-modal hierarchical representations. In: International conference on medical image computing and computer-assisted intervention. pp. 448–458. Springer (2023)
13. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021)
14. Großkopf, E., Bundele, V., Hosseinzadeh, M., Lensch, H.: Histdist: Histopathological diffusion-based stain transfer. arXiv preprint arXiv:2505.06793 (2025)
15. Gurcan, M.N., Boucheron, L.E., Can, A., Madabhushi, A., Rajpoot, N.M., Yener, B.: Histopathological image analysis: A review. IEEE reviews in biomedical engineering **2**, 147–171 (2009)
16. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
17. Li, F., Hu, Z., Chen, W., Kak, A.: Adaptive supervised patchnce loss for learning h&e-to-ihc stain translation with inconsistent groundtruth image pairs. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 632–641. Springer (2023)
18. Lu, M., Wang, T., Xia, Y.: Multi-modal pathological pre-training via masked autoencoders for breast cancer diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 457–466. Springer (2023)
19. Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., et al.: A visual-language foundation model for computational pathology. Nature Medicine **30**(3), 863–874 (2024)
20. Ma, R., Lu, M., Xiang, Z., Pan, Y., Wu, Y., Zeng, Q., Xia, Y.: Paint: Pathology-aware integrated next-scale transformation for virtual immunohistochemistry. arXiv preprint arXiv:2601.16024 (2026)
21. Niazi, M.K.K., Parwani, A.V., Gurcan, M.N.: Digital pathology and artificial intelligence. The lancet oncology **20**(5), e253–e261 (2019)
22. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
23. Özbey, M., Dalmaz, O., Dar, S.U., Bedel, H.A., Özturk, Ş., Güngör, A., Cukur, T.: Unsupervised medical image translation with adversarial diffusion models. IEEE Transactions on Medical Imaging **42**(12), 3524–3539 (2023)
24. Park, T., Efros, A.A., Zhang, R., Zhu, J.Y.: Contrastive learning for unpaired image-to-image translation. In: European conference on computer vision. pp. 319–345. Springer (2020)
25. Razavi, A., Van den Oord, A., Vinyals, O.: Generating diverse high-fidelity images with vq-vae-2. Advances in neural information processing systems **32** (2019)
26. Tian, K., Jiang, Y., Yuan, Z., Peng, B., Wang, L.: Visual autoregressive modeling: Scalable image generation via next-scale prediction. Advances in neural information processing systems **37**, 84839–84865 (2024)

27. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)
28. Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Severson, K., Zimmermann, E., Hall, J., Tenenholtz, N., Fusi, N., et al.: A foundation model for clinical-grade computational pathology and rare cancers detection. Nature medicine **30**(10), 2924–2935 (2024)
29. Wodzinski, M., Marini, N., Atzori, M., Müller, H.: Deeperhistreg: robust whole slide images registration framework. arXiv preprint arXiv:2404.14434 (2024)
30. Wu, M., Goodman, N.: Multimodal generative models for scalable weakly-supervised learning. Advances in neural information processing systems **31** (2018)
31. Zhang, W., Hui, T.H., Tse, P.Y., Hill, F., Lau, C., Li, X.: High-resolution medical image translation via patch alignment-based bidirectional contrastive learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 178–188. Springer (2024)
32. Zheng, B., Ma, N., Tong, S., Xie, S.: Diffusion transformers with representation autoencoders. arXiv preprint arXiv:2510.11690 (2025)
33. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)