PAPER PRESENTATION

HistDiT: A Structure-Aware Latent Conditional Diffusion Model for High-Fidelity Virtual Staining in Histopathology

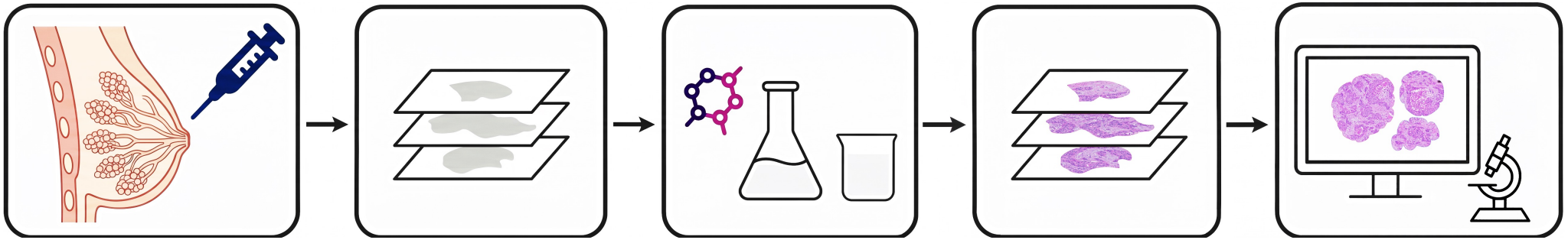
Aasim Bin Saleem · Amr Ahmed · **Ardhendu Behera** · Hafeezullah Amin · Iman Yi Liao · Mahmoud Khattab · Pan Jia Wern · Haslina Makmur

Department of Computer Science, Edge Hill University, United Kingdom · *University of Nottingham Malaysia* · *Cancer Research Malaysia*

Paper: ***arxiv.org/abs/2604.08305*** Code: ***github.com/AasimBinSaleem/HistDiT*** Email: ***aasimr@edgehill.ac.uk***

MOTIVATION · CLINICAL CONTEXT

- Precise treatment for Breast Cancer involve assessment of biomarkers.
- Histological staining serves as the trademark in cancer diagnosis.



❖ **H&E Staining:** Standard, cheap, but lacks biomarker specificity.

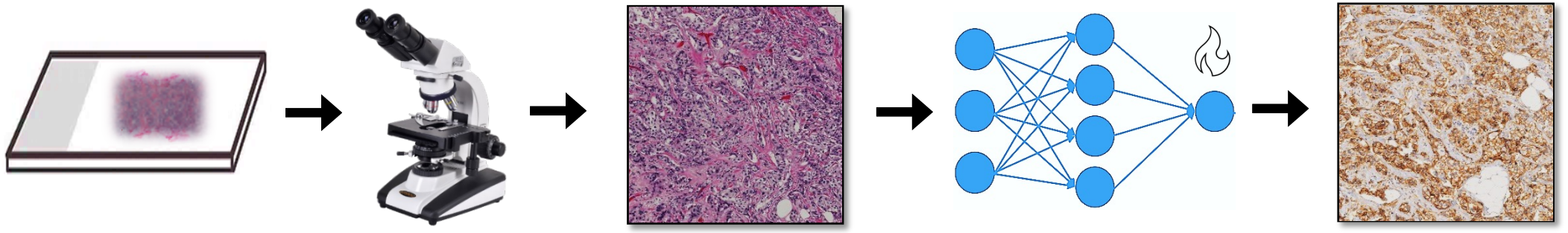
❖ **IHC Staining:** Identifies a variety of immune biomarkers, dictates targeted therapies for early personalized treatments.

IHC is resource-intensive, time-consuming and damages tissue homeostasis.

VIRTUAL STAINING

- Virtual staining uses AI to generate synthetically stained IHC samples.

✓ Provides instant, selective, cost-free and high-contrast diagnostics.

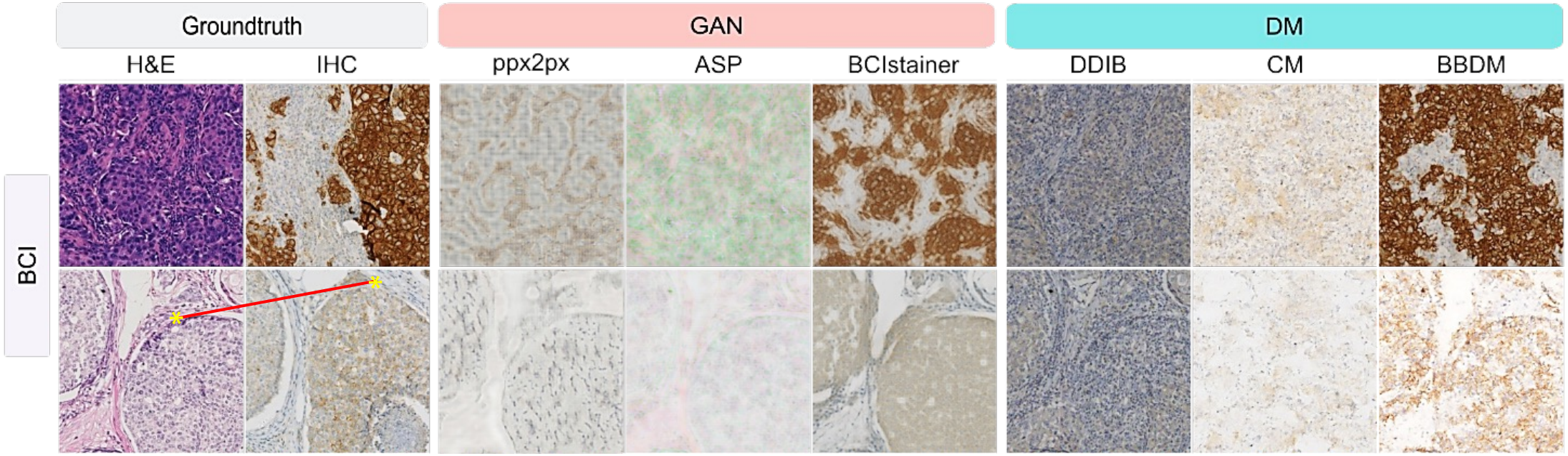


Histological Stain H&E (Type A) | Brightfield Imaging | Stained Image **H&E (Type A)** | Deep Learning Model (Trained) | Virtually Stained **IHC (Type B)**

GOAL

H&E stained to IHC stained image translation for breast cancer prognosis.

PRIOR WORK

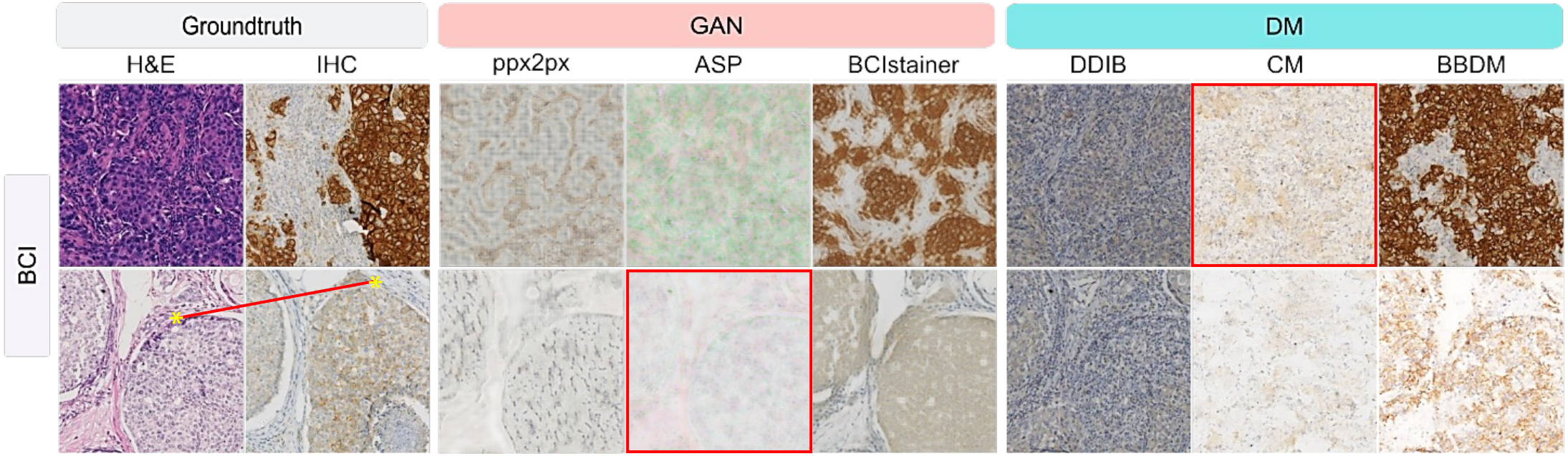


Examples for GANs and Diffusion models, trained on BCI Dataset [6]

⚠ LIMITATIONS:

- **GANs:** Difficult to train – suffer from mode collapse issues; Wash out high expression regions.
- **Standard Diffusion (U-Nets):** Structure and staining trade-offs.
  Struggle with global context – hallucinate cells and misalign structures in complex tissues.

PRIOR WORK

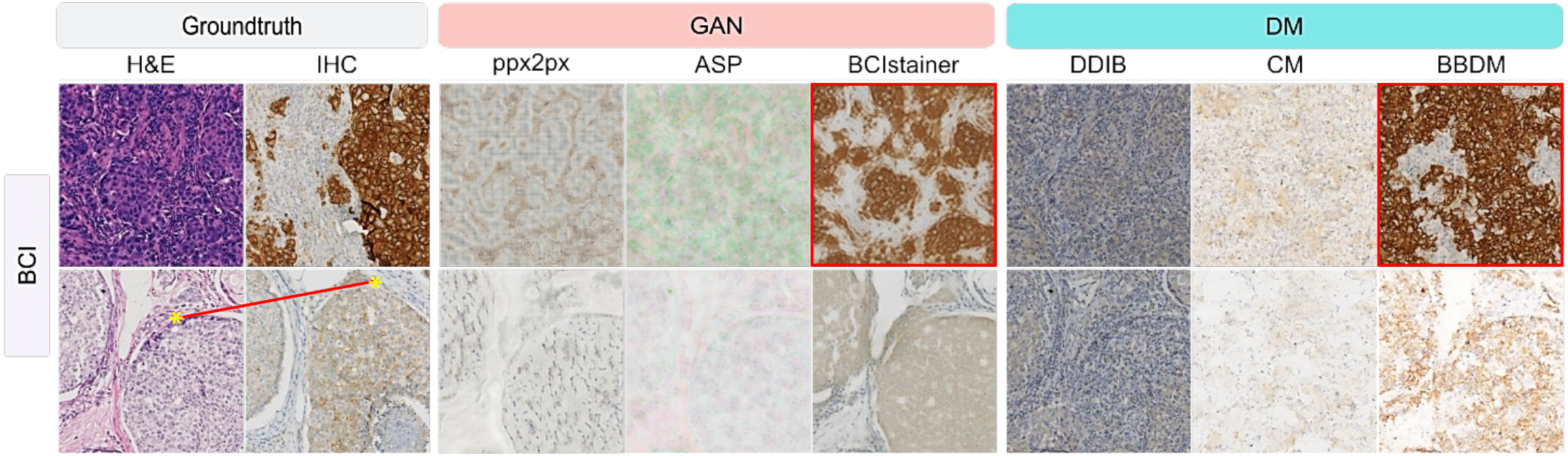


Examples for GANs and Diffusion models, trained on BCI Dataset [6]

⚠ LIMITATIONS:

- **GANs:** Difficult to train – suffer from mode collapse issues; Wash out high expression regions.
- **Standard Diffusion (U-Nets):** Structure and staining trade-offs.
  Struggle with global context – hallucinate cells and misalign structures in complex tissues.

PRIOR WORK

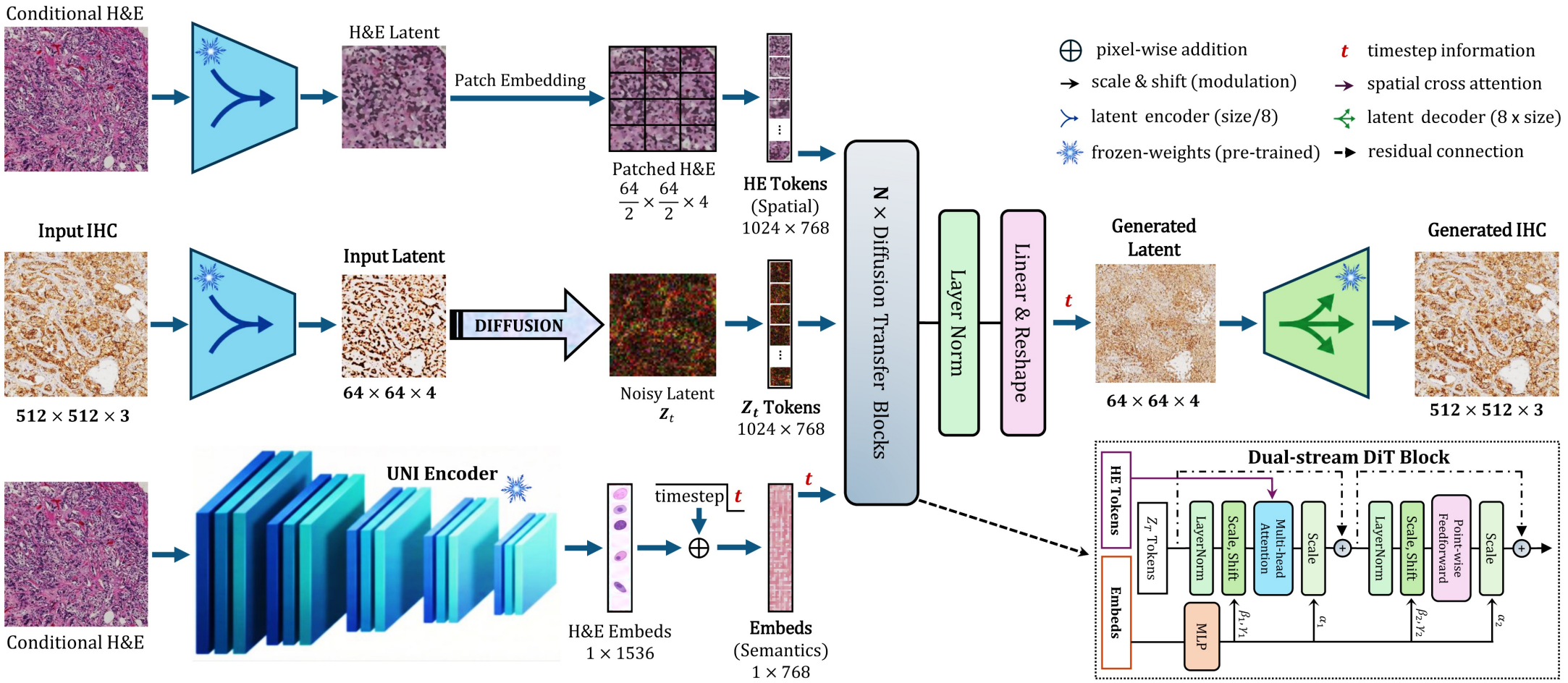


Examples for GANs and Diffusion models, trained on BCI Dataset [6]

⚠ LIMITATIONS:

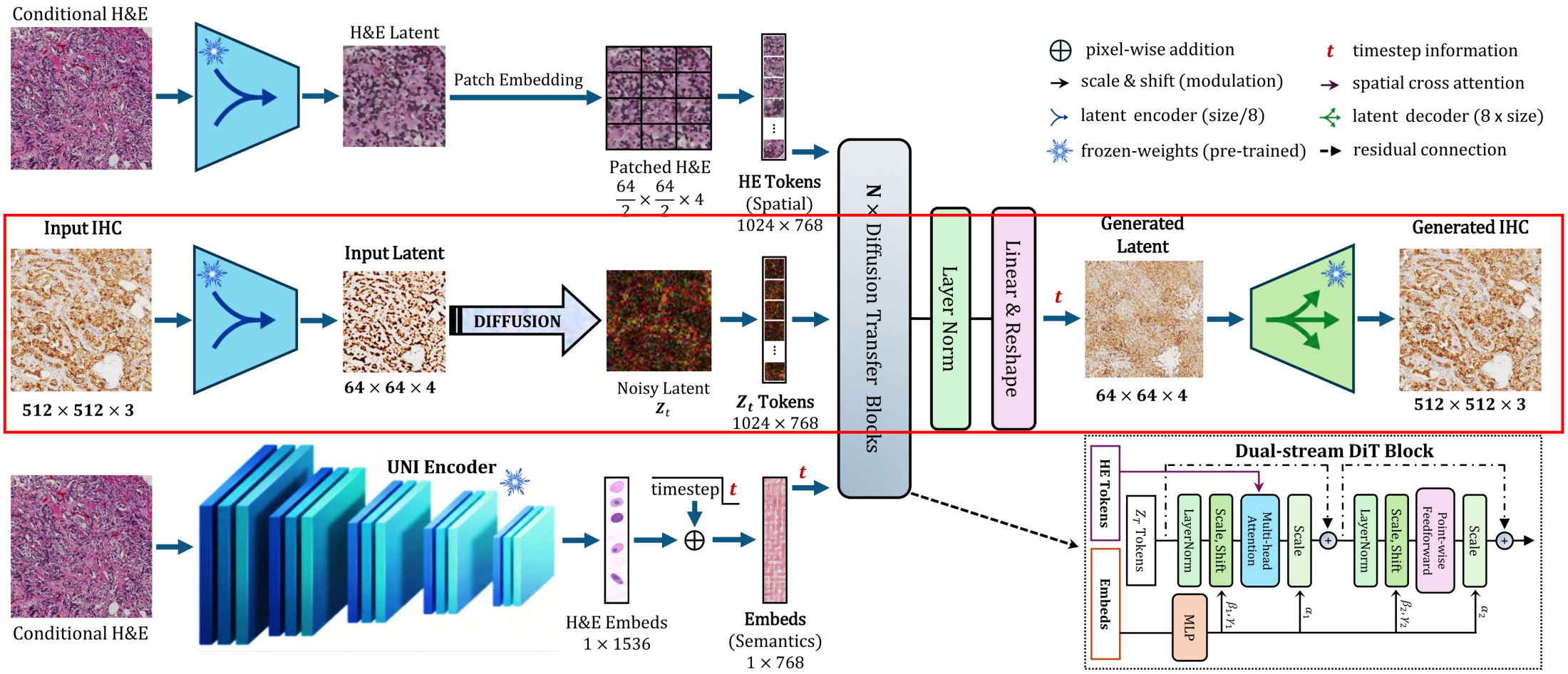
- **GANs:** Difficult to train – suffer from mode collapse issues; Wash out high expression regions.
- **Standard Diffusion (U-Nets):** Structure and staining trade-offs.
  Struggle with global context – hallucinate cells and misalign structures in complex tissues.

PROPOSED VIRTUAL STAINING FRAMEWORK



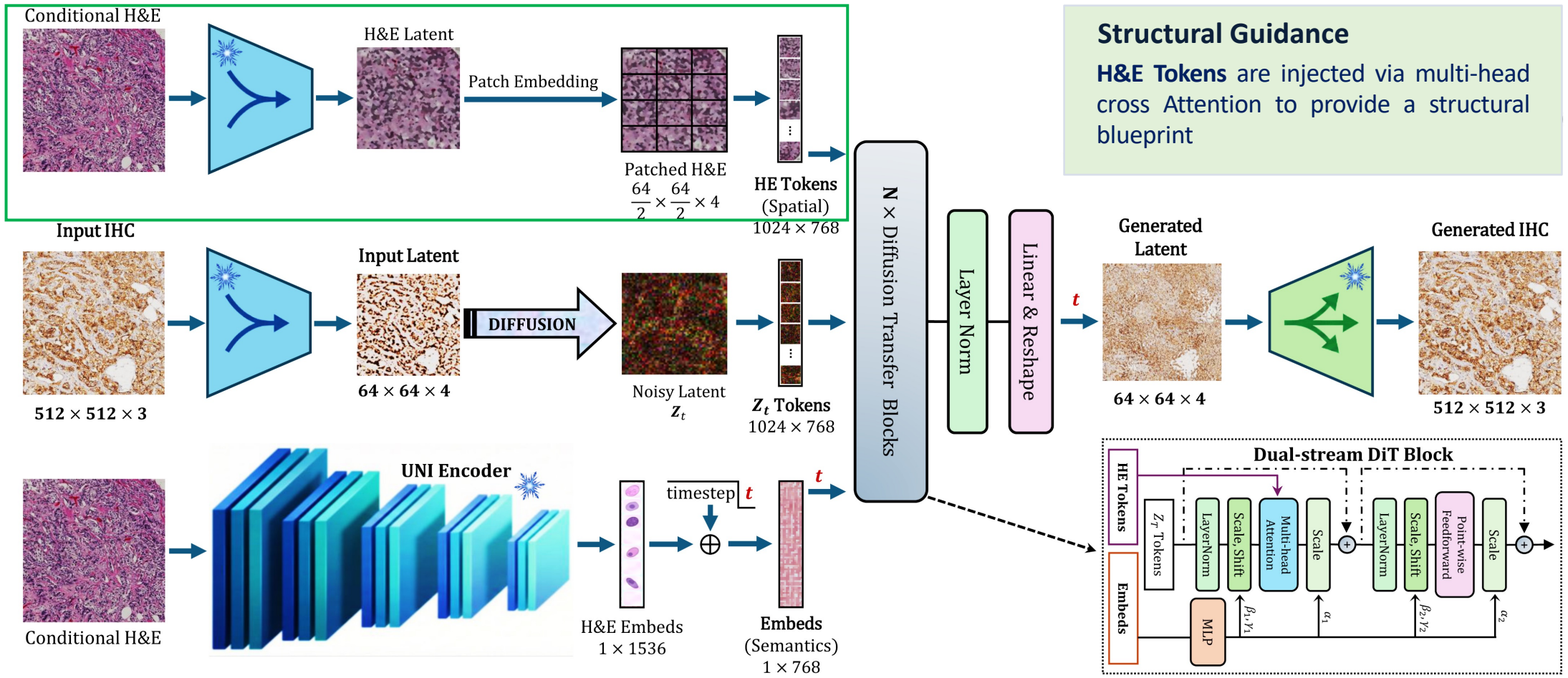
Training objective: $\mathcal{L}\_Train = \alpha \cdot \mathcal{L}\_MSE + (1 - \alpha) \cdot \mathcal{L}\_L1$ · **Noise Schedular: Standard DDPM** *(epsilon prediction)*

PROPOSED VIRTUAL STAINING FRAMEWORK



Training objective: $\mathcal{L}\_Train = \alpha \cdot \mathcal{L}\_MSE + (1 - \alpha) \cdot \mathcal{L}\_L1$ · **Noise Schedular: Standard DDPM** *(epsilon prediction)*

PROPOSED VIRTUAL STAINING FRAMEWORK

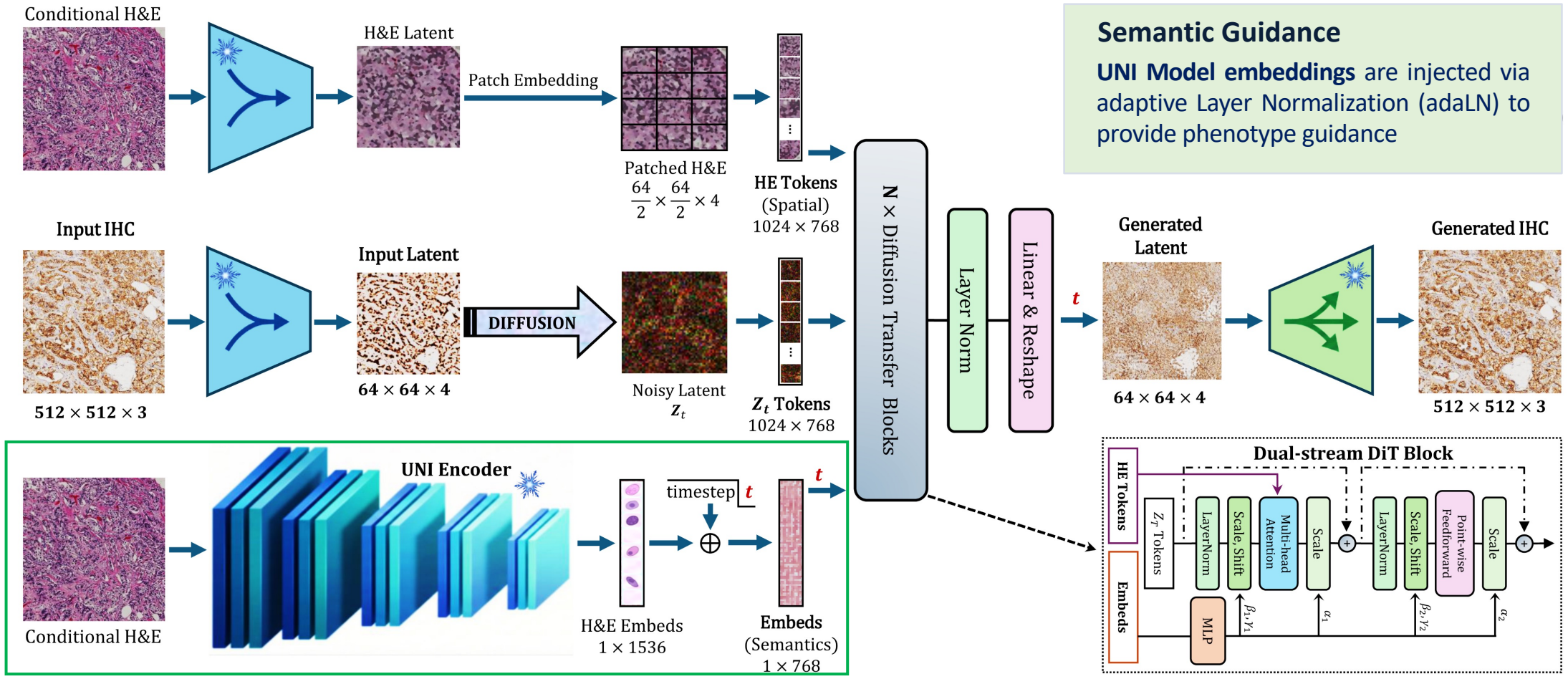


Structural Guidance

H&E Tokens are injected via multi-head cross Attention to provide a structural blueprint

Training objective: $\mathcal{L}\_Train = \alpha \cdot \mathcal{L}\_MSE + (1 - \alpha) \cdot \mathcal{L}\_L1$ · **Noise Schedular:** **Standard DDPM** *(epsilon prediction)*

PROPOSED VIRTUAL STAINING FRAMEWORK

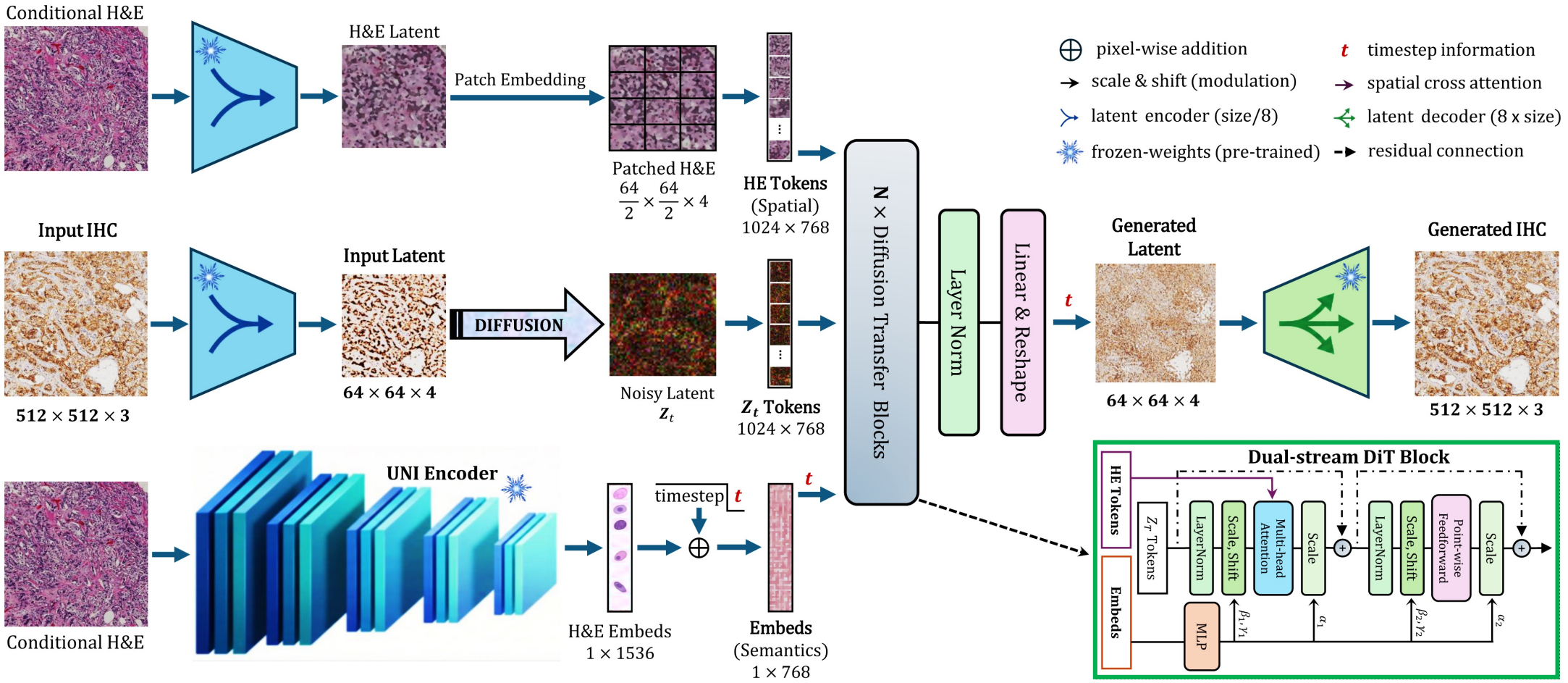


Semantic Guidance

UNI Model embeddings are injected via adaptive Layer Normalization (adaLN) to provide phenotype guidance

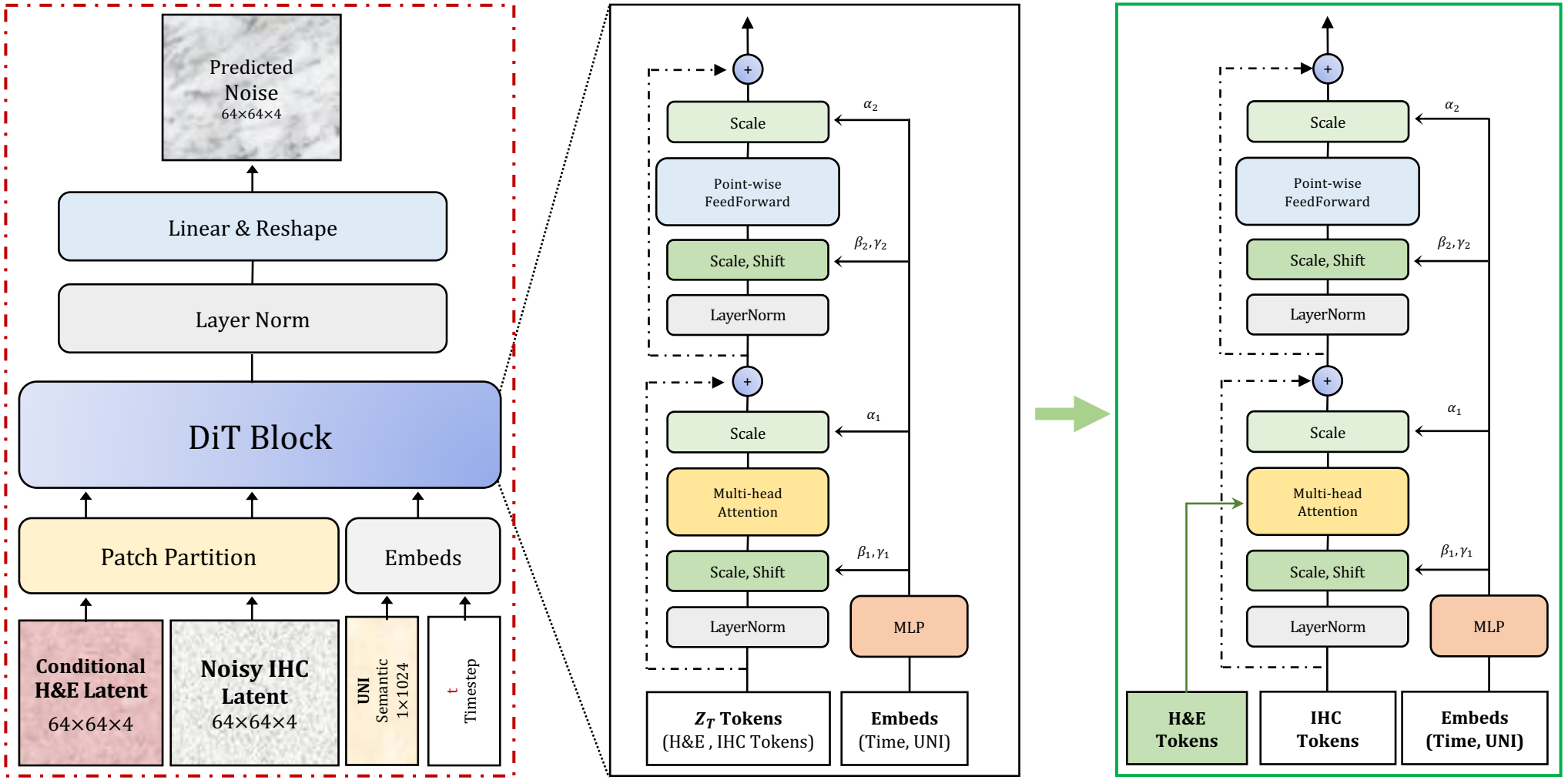
Training objective: $\mathcal{L}\_Train = \alpha \cdot \mathcal{L}\_MSE + (1 - \alpha) \cdot \mathcal{L}\_L1$ · **Noise Schedular:** **Standard DDPM** *(epsilon prediction)*

PROPOSED VIRTUAL STAINING FRAMEWORK

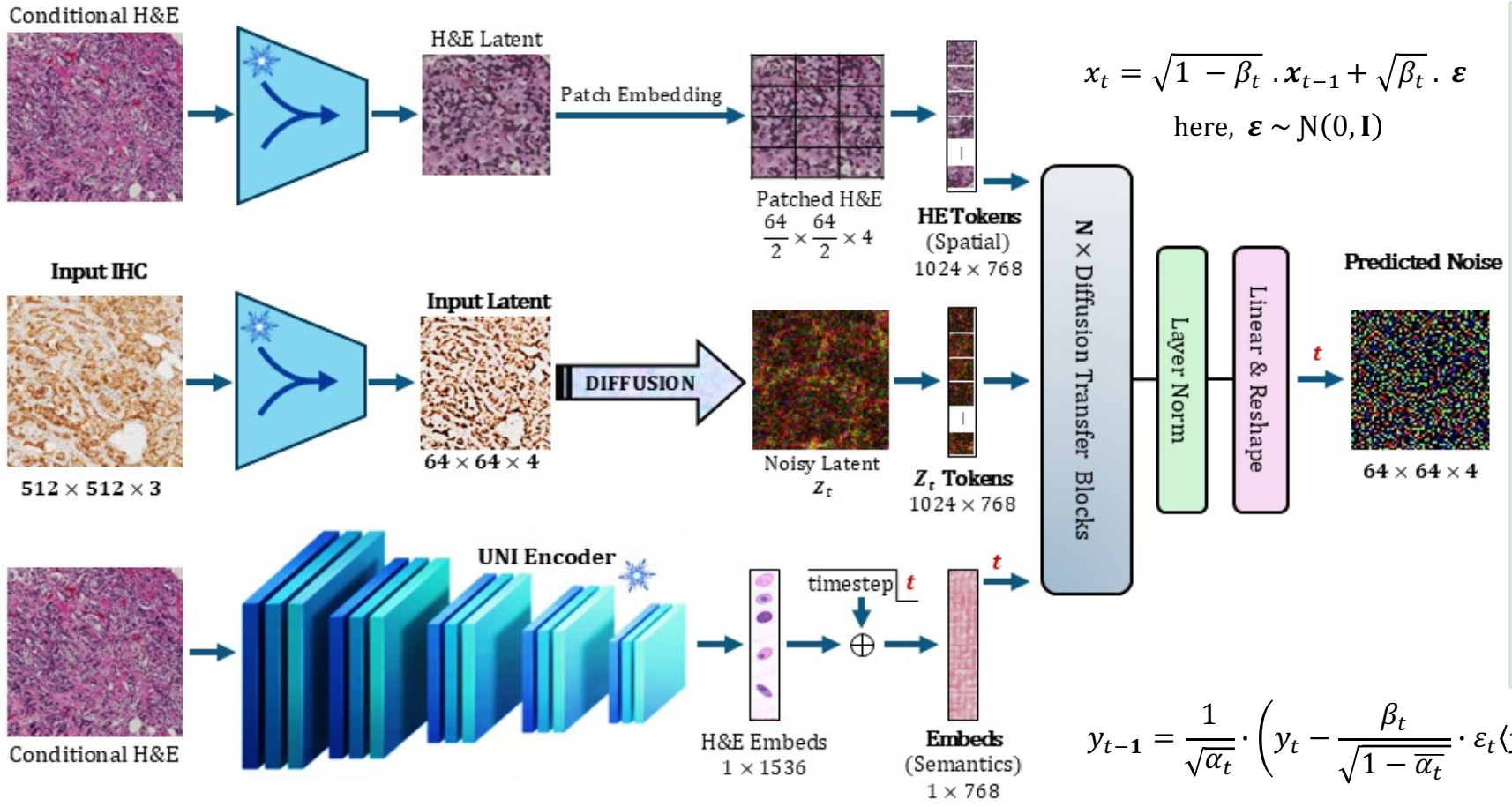


Training objective: $\mathcal{L}\_Train = \alpha \cdot \mathcal{L}\_MSE + (1 - \alpha) \cdot \mathcal{L}\_L1$ · **Noise Schedular: Standard DDPM** *(epsilon prediction)*

DUAL-STREAM DIT BLOCK



TRAINING PIPELINE



$$x_t = \sqrt{1-\beta_t}\,.\,\boldsymbol{x}_{t-1} + \sqrt{\beta_t}\,.\,\boldsymbol{\varepsilon}$$

here, $\boldsymbol{\varepsilon} \sim \mathrm{N}(0, \mathbf{I})$

Multi-Objective Loss

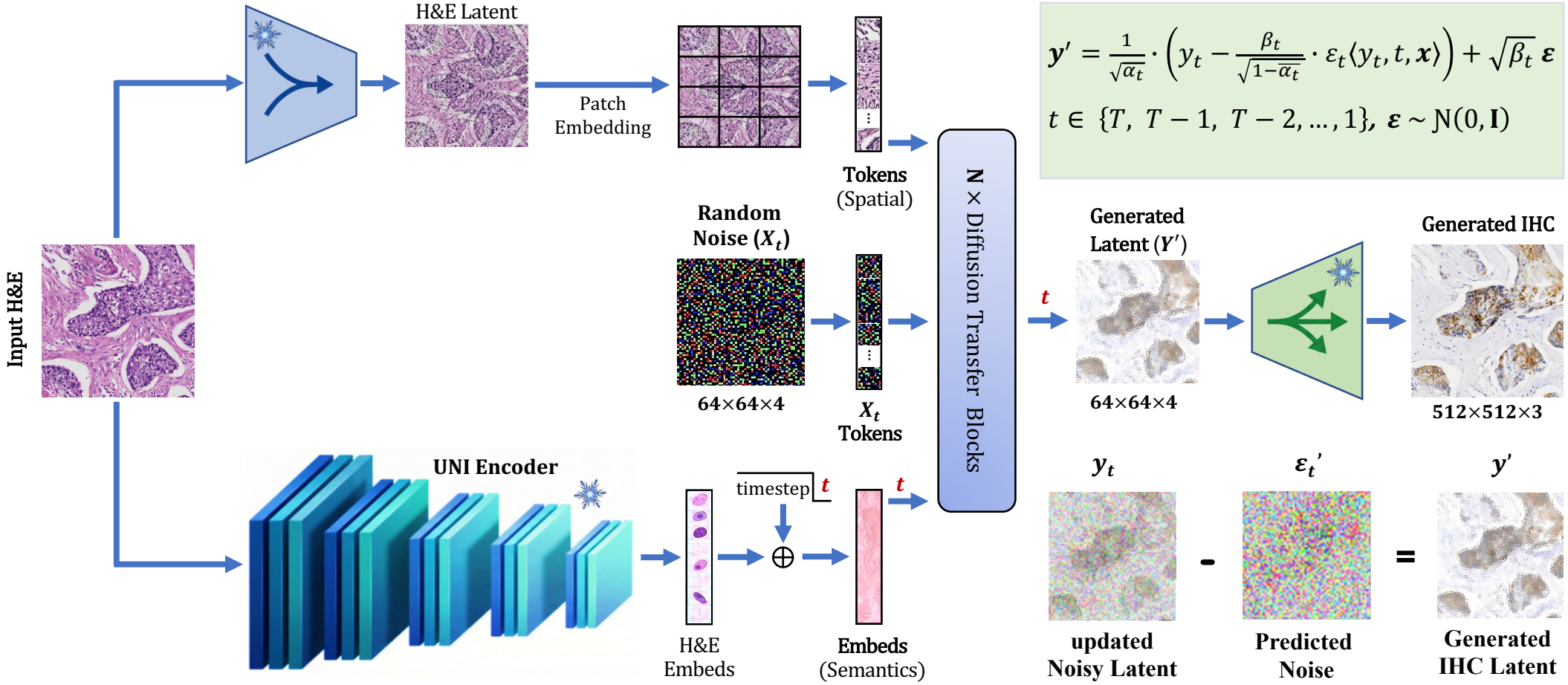
Combining an auxiliary $\boldsymbol{L_1}$ term with the standard MSE

$$L_{Total} = \mathbf{0.7}\times L_{MSE} + \mathbf{0.3}\times L_{L_1}$$

$$y_{t-1} = \frac{1}{\sqrt{\alpha_t}} \cdot \left(y_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha_t}}} \cdot \varepsilon_t \langle y_t, t, \boldsymbol{x} \rangle \right) + \sqrt{\beta_t}\boldsymbol{\varepsilon}$$

Training objective: $\mathcal{L}\_\text{Train} = \alpha \cdot \mathcal{L}\_\text{MSE} + (1 - \alpha) \cdot \mathcal{L}\_\text{L1}$ · **Noise Schedular:** Standard DDPM *(epsilon prediction)*

THE INFERENCE



$$y' = \frac{1}{\sqrt{\alpha_t}} \cdot \left(y_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \cdot \varepsilon_t \langle y_t, t, x \rangle \right) + \sqrt{\beta_t}\, \varepsilon$$

$$t \in \{T,\ T-1,\ T-2, \ldots, 1\},\ \varepsilon \sim \mathrm{N}(0, \mathbf{I})$$

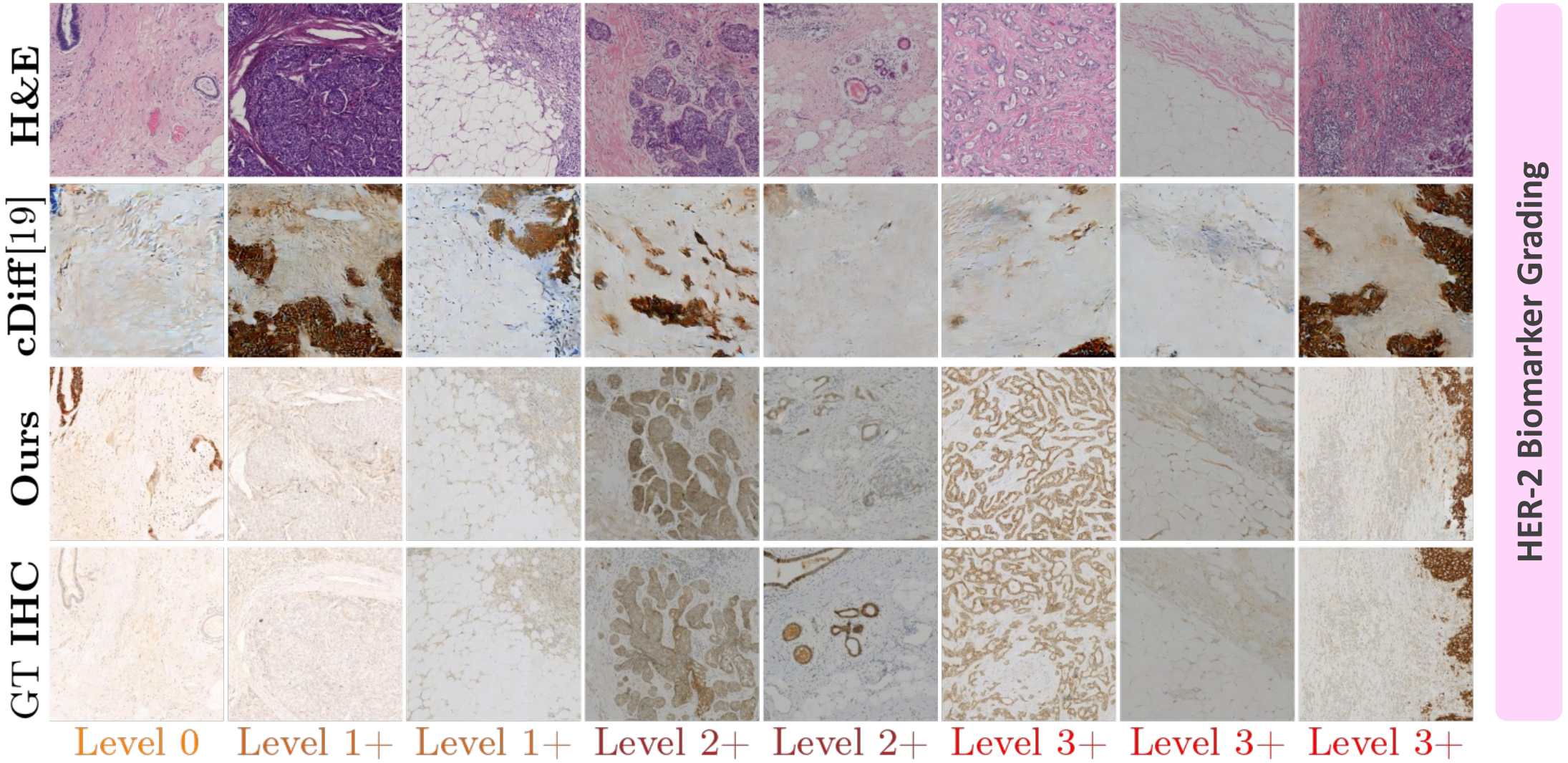
Noise Schedular: Standard DDPM *(scaled linear)* · **Prediction Type:** Epsilon · **Sampling Steps:** 1000

QUANTITATIVE COMARISON ON BCI DATASET

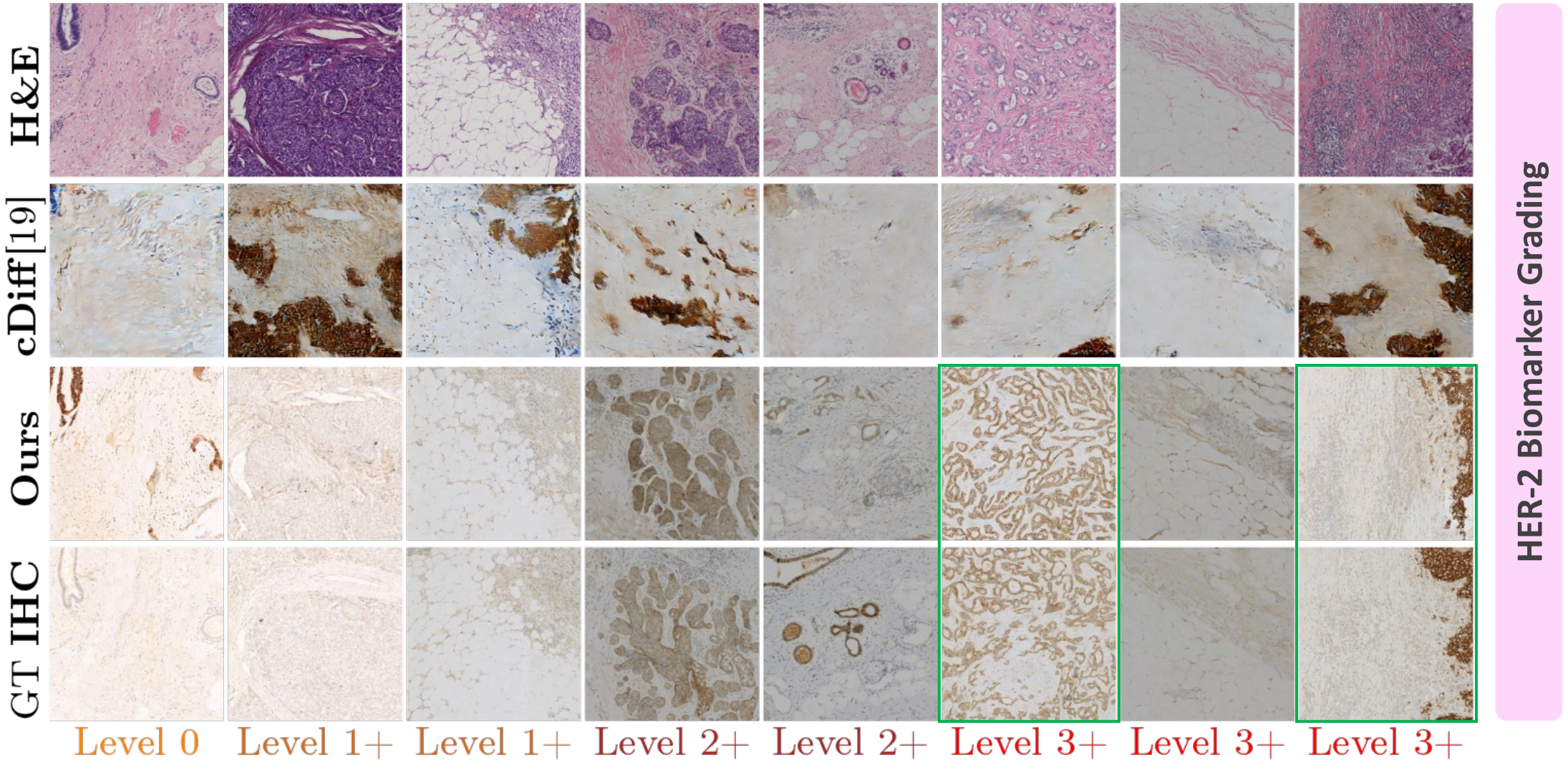
Method/Model	MSE ↓	PSNR(dB) ↑	SSIM ↑	SCM ↑	LPIPS ↓	FID ↓	
Cycle GAN [4] ɕ	1892.00	15.98	0.372	0.447	0.624	188.33	GANs
Pix2Pix [5] ɕ	1560.26	17.08	0.303	0.391	0.769	<u>160.39</u>	GANs
Pyramid Pix2Pix [6] ɕ	<u>1348.54</u>	19.61	0.397	0.473	<u>0.466</u>	167.40	GANs
ASP [7] ɕ	---	---	0.5032	---	---	---	GANs
Syn Diff [8] ʑ	---	14.28±2.52	0.32±0.1	---	---	---	Diffusion
PST-Diff [9] ʑ	---	16.75±4.20	0.38±0.1	---	---	---	Diffusion
Conditional Diff. [19] ʑ	3011.36	15.76	0.405	<u>0.494</u>	0.591	194.82	Diffusion
Star-Diff (No Visual) ʑ	---	<u>21.30±0.01</u>	0.5301	---	---	---	Diffusion
HistDiST [21] ʑ	---	---	<u>0.4693</u>	---	---	---	Diffusion
Proposed [HistDiT] ʑ	**891.53**	**21.43**	0.4769	**0.540**	**0.412**	**49.15**	OUR
%age Improvement	33.89%	0.66%	5.2%	9.31%	11.59%	69.4%	

Structural Correlation Metric **(SCM)** mitigates mathematical bias in standard SSIM towards brightfield microscopy.

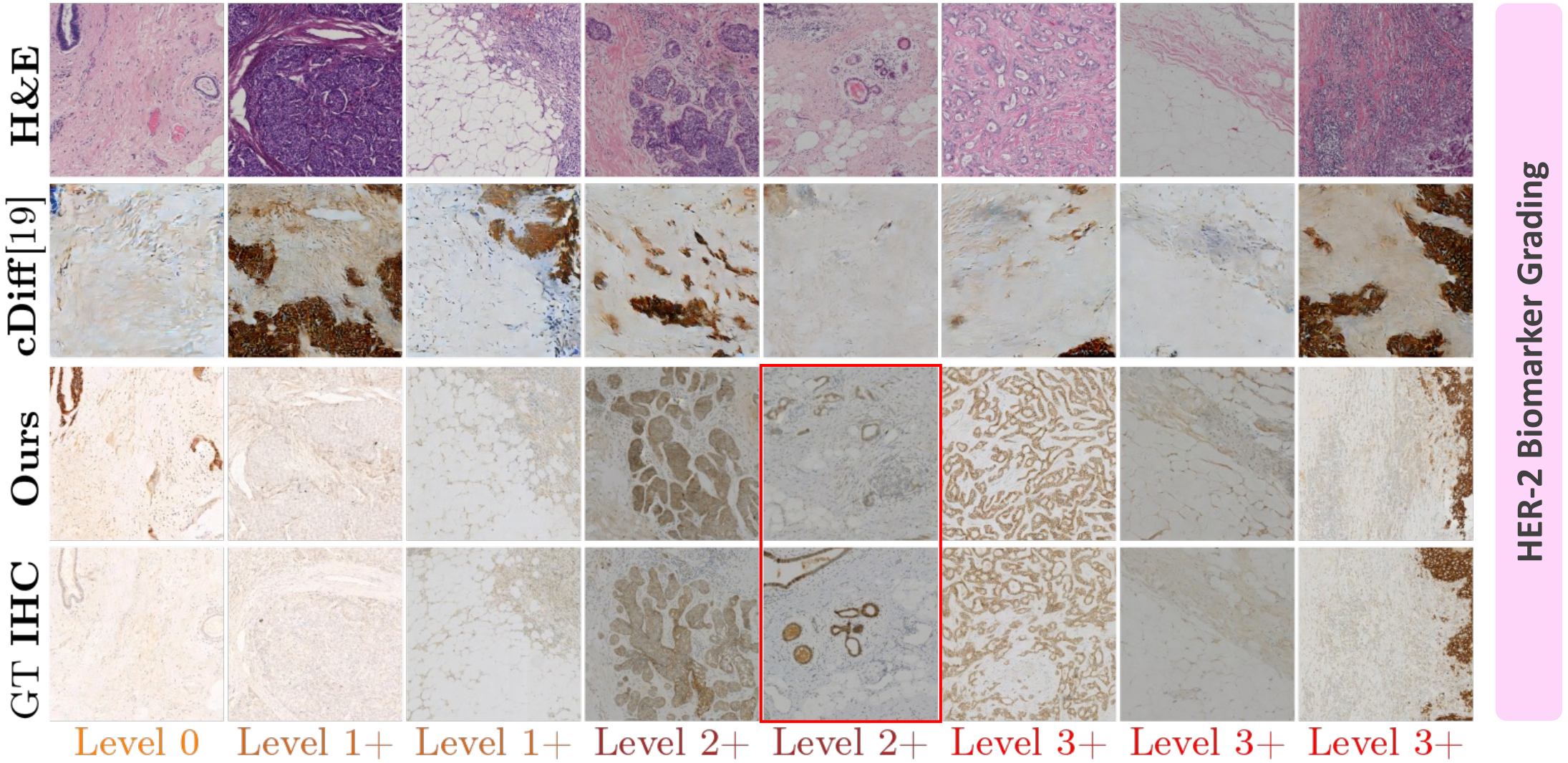
VISUAL COMARISON ON THE BCI DATASET



VISUAL COMARISON ON THE BCI DATASET



VISUAL COMARISON ON THE BCI DATASET

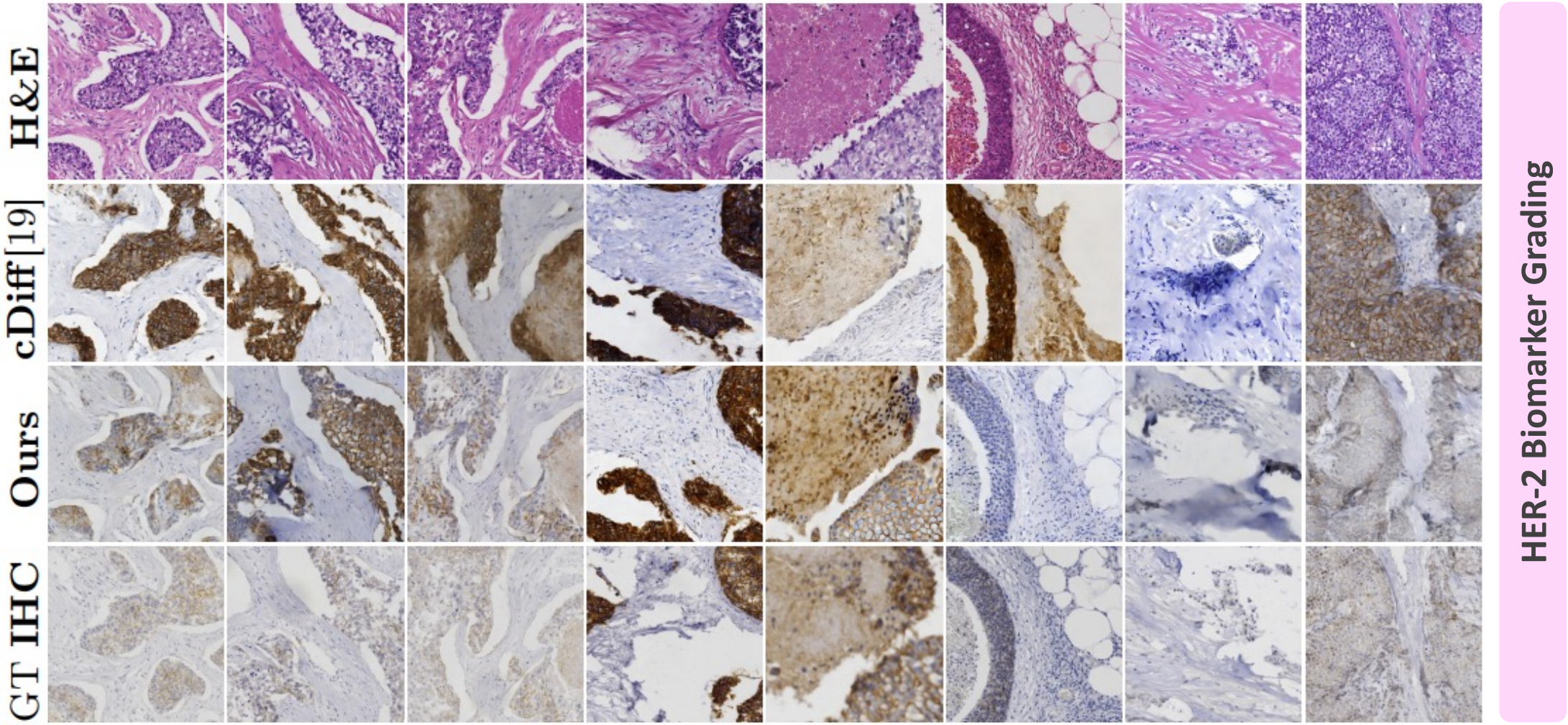


QUANTITATIVE COMARISON ON MIST DATASET

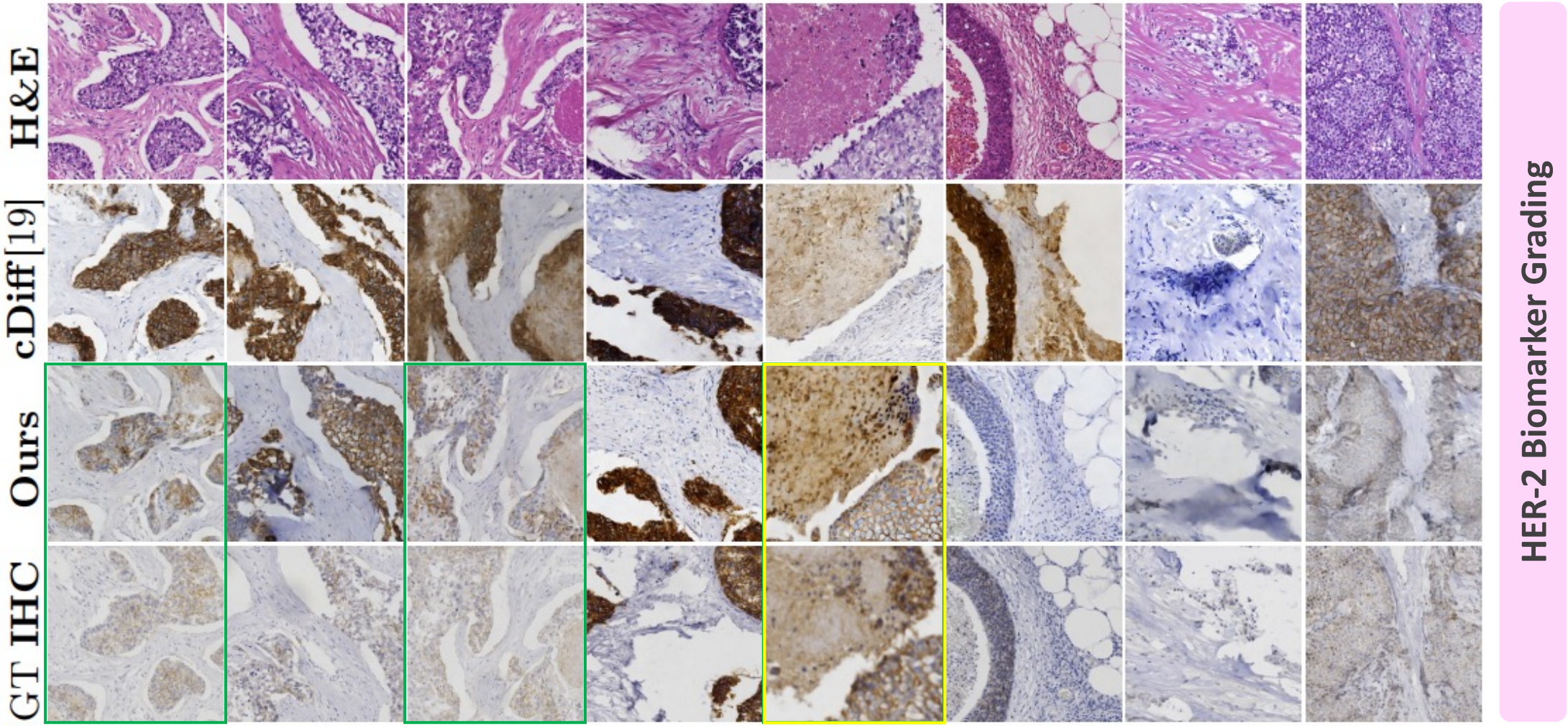
Method/Model	MSE ↓	PSNR(dB) ↑	SSIM ↑	SCM ↑	LPIPS ↓	FID ↓	
Cycle GAN [4] ᴳ	4190.52	11.91	0.174	0.209	0.553	125.7	GANs
Pix2Pix [5] ᴳ	3485.20	12.74	0.150	0.194	0.614	128.1	GANs
Pyramid Pix2Pix [6] ᴳ	<u>2986.44</u>	<u>14.24</u>	0.165	0.214	<u>0.543</u>	107.4	GANs
ASP [7] ᴳ	6252.98	11.13	0.168	<u>0.254</u>	0.564	82.6	GANs
Conditional Diff. [19] ᶻ	---	---	0.1810	---	---	---	DMs
HistDiST [21] ᶻ	---	---	0.1945	---	---	**54.28**	DMs
PixCell (no LoRA) [38] ᶻ	---	---	0.1880	---	---	67.68	DMs
Proposed [HistDiT] ᶻ	3396.8	**14.26**	**0.211**	**0.302**	**0.489**	59.30	OUR
%age Improvement	13.74%	0.14%	2.5%	18.9%	9.94%	9.25%	

Standard **SSIM** has a mathematical bias towards brightfield microscopy, where luminance masks structural errors. Therefore, we suggest the Structural Correlation Metric (**SCM**) for more honest comparison.

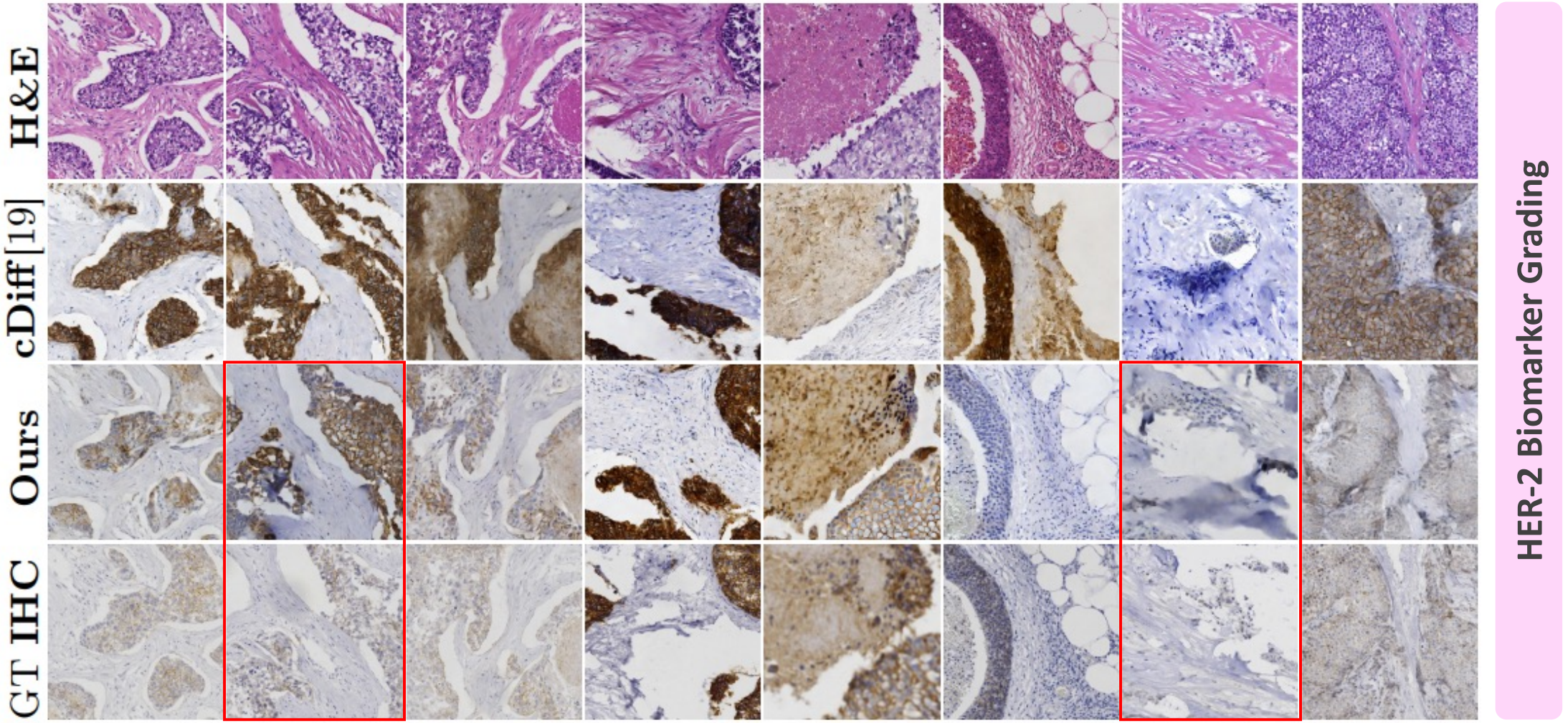
VISUAL COMARISON ON MIST DATASET



VISUAL COMARISON ON MIST DATASET



VISUAL COMARISON ON MIST DATASET



ABLATION STUDIES

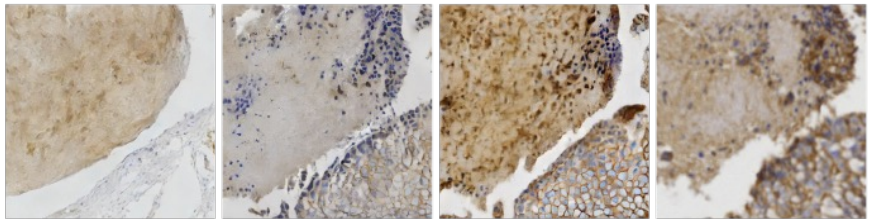
Impact of Dual-stream Conditioning

Model Configuration	PSNR(dB)↑	SSIM↑	SCM↑	LPIPS↓	FID↓
Pyramid Pix2Pix [6]$^{\rho}$	19.61	0.397	0.473	0.466	167.40
HistDiT (Semantics only)^{z}	16.39	0.423	0.492	0.553	81.51
HistDiT (Spatial only with Cross-Attention)^{z}	18.58	0.416	**0.560**	0.452	62.38
HistDiT (Spatial [Concatenation], Semantics)^{z}	20.63	0.449	0.521	0.438	56.80
HistDiT (Spatial [Cross-Attention], Semantics)^{z}	**21.43**	**0.477**	0.540	**0.412**	**49.15**

Table Comparison for different architectural choices on BCI Dataset. Best are **bold**.

Impact of Multi-objective Training Loss

Method/Approach	PSNR↑	SSIM↑	FID↓
MSE only	16.86	0.438	67.88
L_1 only	19.44	0.450	93.31
$0.9MSE + 0.1L_1$	20.44	0.443	56.18
$0.7MSE + 0.3L_1$	**21.43**	**0.477**	**49.15**



MSE only L1 only MSE+L1 Real IHC

Fig. Ablation Study on Objective Functions: (*Left*) Quantitative comparison of loss components on the BCI dataset. $MSE+L_1$ maintains high metric scores. (*Right*) Visual samples demonstrating that the combined objective produces sharper structures (column 3) compared to the smoothing artifacts seen in MSE only (column 1).

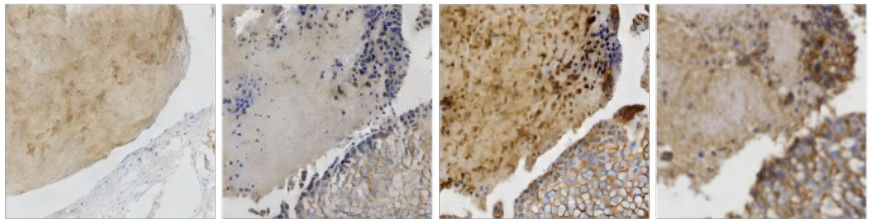
ABLATION STUDIES

Model Configuration	PSNR(dB)↑	SSIM↑	SCM↑	LPIPS↓	FID↓
Pyramid Pix2Pix [6]$^{\rho}$	19.61	0.397	0.473	0.466	167.40
HistDiT (Semantics only)^{z}	16.39	0.423	0.492	0.553	81.51
HistDiT (Spatial only with Cross-Attention)^{z}	18.58	0.416	**0.560**	0.452	62.38
HistDiT (Spatial [Concatenation], Semantics)^{z}	20.63	0.449	0.521	0.438	56.80
HistDiT (Spatial [Cross-Attention], Semantics)^{z}	**21.43**	**0.477**	0.540	**0.412**	**49.15**

Table Comparison for different architectural choices on BCI Dataset. Best are **bold**.

Impact of Dual-stream Conditioning

Method/Approach	PSNR↑	SSIM↑	FID↓
MSE only	16.86	0.438	67.88
L_1 only	19.44	0.450	93.31
$0.9MSE + 0.1L_1$	20.44	0.443	56.18
$0.7MSE + 0.3L_1$	**21.43**	**0.477**	**49.15**



MSE only **L1 only** **MSE+L1** **Real IHC**

Impact of Multi-objective Training Loss

Fig. Ablation Study on Objective Functions: (*Left*) Quantitative comparison of loss components on the BCI dataset. $MSE+L_1$ maintains high metric scores. (*Right*) Visual samples demonstrating that the combined objective produces sharper structures (column 3) compared to the smoothing artifacts seen in MSE only (column 1).

ABLATION STUDIES

Impact of Dual-stream Conditioning

Model Configuration	PSNR(dB)↑	SSIM↑	SCM↑	LPIPS↓	FID↓
Pyramid Pix2Pix [6]$^{\rho}$	19.61	0.397	0.473	0.466	167.40
HistDiT (Semantics only)^{z}	16.39	0.423	0.492	0.553	81.51
HistDiT (Spatial only with Cross-Attention)^{z}	18.58	0.416	**0.560**	0.452	62.38
HistDiT (Spatial [Concatenation], Semantics)^{z}	20.63	0.449	0.521	0.438	56.80
HistDiT (Spatial [Cross-Attention], Semantics)^{z}	**21.43**	**0.477**	0.540	**0.412**	**49.15**

Table Comparison for different architectural choices on BCI Dataset. Best are **bold**.

Impact of Multi-objective Training Loss

Method/Approach	PSNR↑	SSIM↑	FID↓
MSE only	16.86	0.438	67.88
L_1 only	19.44	0.450	93.31
$0.9MSE + 0.1L_1$	20.44	0.443	56.18
$0.7MSE + 0.3L_1$	**21.43**	**0.477**	**49.15**

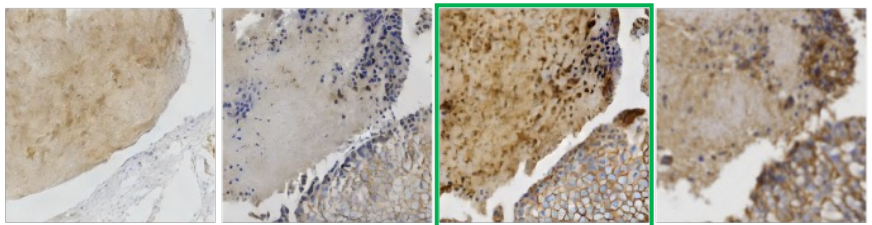


Fig. Ablation Study on Objective Functions: (*Left*) Quantitative comparison of loss components on the BCI dataset. $MSE+L_1$ maintains high metric scores. (*Right*) Visual samples demonstrating that the combined objective produces sharper structures (column 3) compared to the smoothing artifacts seen in MSE only (column 1).

VALIDATION & FUTURE DIRECTIONS

❖ Experts from CRMY confirmed the staining fidelity of our generated samples via blind qualitative assessment: Indistinguishable from real IHC.

LIMITATIONS & FUTURE WORK

READ MORE

- Computational overhead as main limitation.
- Domain gap for using latent encoders that offer lossy compression.
- Detailed HER-2 grading analysis, including Cohen's Kappa to be obtained.
- Stain normalization for cross-dataset evaluation.

Paper: ***arxiv.org/abs/2604.08305*** Code: ***github.com/AasimBinSaleem/HistDiT*** Email: ***aasimr@edgehill.ac.uk***

REFERENCES

1. B. Bai et al., “Deep learning-enabled virtual histological staining of biological samples,” Light: Science & Applications, vol. 12, no. 1, p. 57, 2023.
2. K. de Haan et al., “Deep learning-based transformation of H&E-stained tissues into special stains,” Nature communications 12(1), 4884. (2021)
3. J. J. Levy et al., “A large-scale internal validation study of unsupervised virtual trichrome staining technologies on nonalcoholic steatohepatitis liver biopsies,” Mod Pathol, vol. 34, pp. 808–822. (2021)
4. S. Liu et al., “Unpaired stain transfer using pathology-consistent constrained generative adversarial networks,” IEEE Transactions on Medical Imaging, vol. 40, no. 8, pp. 1977-1989. (2021)
5. C. Mercan et al., “Virtual staining for mitosis detection in breast histopathology,” in 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 1770-1774, IEEE. (2020)
6. P. Klöckner, J. Teixeira, D. Montezuma, J. Cardoso, H. Horlingsand S. P. Oliveira, “HER2match dataset”. Zenodo (2025)
7. Zhang et al., “Gan-based virtual re-staining: a promising solution for whole slide image analysis,” arXiv preprint arXiv:1901.04059. (2019)
8. Liu, S., et al.: “BCI: Breast cancer immunohistochemical image generation through pyramid pix2pix.” In: CVPR W, pp. 1815–1824. IEEE (2022)
9. He, Y., et al.: “PST-Diff: Achieving high-consistency stain transfer by diffusion models with pathological and structural constraints.” IEEE TMI (2024)
10. Er, X., et al.: Conditional diffusion-based virtual staining. In: ICPR, pp. 193–207 Springer, Cham (2024)
11. Rombach, R., et al.: High-resolution image synthesis with latent diffusion models. In: CVPR, pp. 10684–10695. IEEE (2022)
12. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV, pp.4195–4205. IEEE (2023)
13. Chen, R.J., et al.: Towards a general-purpose foundation model for computational pathology. Nature Medicine 30(3), 850–862 (2024)

READ MORE