

# Supplementary Materials: Multi-modal Denoising Diffusion Pre-training for Whole-Slide Image Classification

Anonymous Authors

## 1 INVESTIGATION OF LOSS WEIGHTS

We investigate how different values of loss weights in the loss function  $\mathcal{L}(X^{he}, X^{ihc}, \rho) = \lambda_1 \mathcal{L}_{rec}^{he \rightarrow ihc} + \lambda_2 \mathcal{L}_{rec}^{ihc \rightarrow ihc} + \lambda_3 \mathcal{L}_{con}$  affect the representation learning capability of our framework. All experiments are conducted on the Camelyon16 dataset using CLAM-SB [2] as the classifier. In Table 1 (a), we examine the values of  $\lambda_1$  while setting  $\lambda_2$  to 1 and  $\lambda_3$  to 0.1. The results show that increasing  $\lambda_2$  from 1 to 10 results in a 2.3% increase in AUC. However, further increasing  $\lambda_1$  to 20 led to a 3.2% decrease in AUC. In Table 1 (b), with  $\lambda_1$  as 10 and  $\lambda_3$  as 0.1, the results reveal that as  $\lambda_2$  increased from 0.5 to 10. The AUC initially increases by 1.1% when  $\lambda_2$  reaches 1, and then gradually decreases. In Table 1 (c), the results demonstrate that increasing  $\lambda_3$  from 0.1 to 0.3 has little impact on the classification AUC. Thus, our framework is less sensitive to  $\lambda_3$ , as  $\lambda_3$  grows from 0.1 to 0.3.

$\lambda_2=1, \lambda_3=0.1$	AUC			
	1	5	10	20
$\lambda_1$	0.845 <sub>0.02</sub>	0.851 <sub>0.04</sub>	<b>0.868</b> <sub>0.04</sub>	0.836 <sub>0.06</sub>

(a)

$\lambda_1=10, \lambda_3=0.1$	AUC			
	0.5	1	5	10
$\lambda_2$	0.857 <sub>0.06</sub>	<b>0.868</b> <sub>0.04</sub>	0.854 <sub>0.04</sub>	0.838 <sub>0.07</sub>

(b)

$\lambda_1=10, \lambda_2=1$	AUC				
	0.1	0.3	0.5	0.7	1.0
$\lambda_3$	<b>0.868</b> <sub>0.04</sub>	0.866 <sub>0.03</sub>	0.858 <sub>0.02</sub>	0.861 <sub>0.02</sub>	0.843 <sub>0.04</sub>

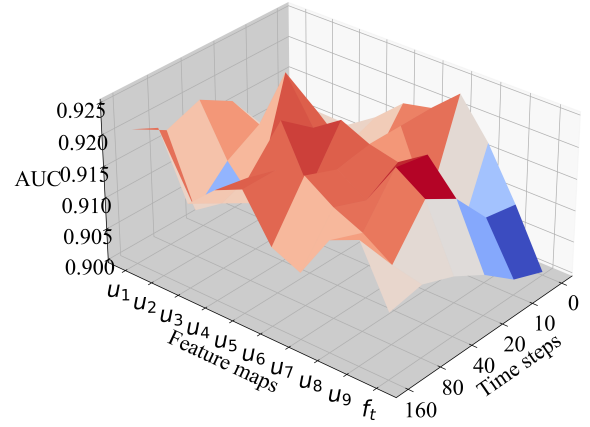
(c)

**Table 1: Loss weights investigation on the Camelyon16 dataset using CLAM-SB [2] as the WSI classifier.**

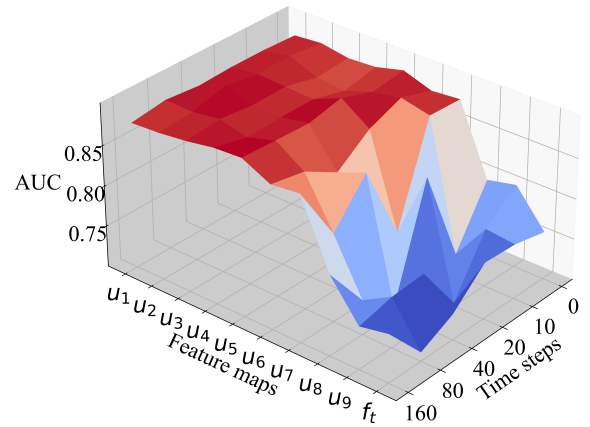
## 2 MORE INVESTIGATION OF FEATURE MAP AND TIME STEP

We investigate the choices of feature maps and time steps on TCGA-NSCLC dataset and TCGA-COAD dataset. In Figure 1 and Figure 2, the height of each point on the 3D mesh represents the AUC value of a WSI classifier trained based on a specific feature map and time step  $t$ . For TCGA-NSCLC dataset, We observe that the feature maps from intermediate decoder layers ( $u_3 - u_7$ ) are more likely to achieve better performance. Regarding the TCGA-COAD dataset, the results in Figure 2 consistently demonstrate that the feature maps from shallow layers ( $u_1 - u_6$ ) outperform feature maps from deep layers. As the above results show, when using our framework on a new dataset, it is suggested to adopt the feature maps from intermediate layers (such as  $u_3 - u_6$ ) or try different feature maps on

the validation set. Regarding the choices of time steps, we find that they have little influence on the representation quality for the WSI classification task on both datasets.



**Figure 1: Hyper-parameter investigation of different time steps and different feature maps on the TCGA-NSCLC dataset using CLAM-SB[2] as the classifier.**



**Figure 2: Hyper-parameter investigation of different time steps and different feature maps on the TCGA-COAD dataset using CLAM-SB[2] as the classifier.**

### 3 MORE VISUAL RESULTS

#### 3.1 Re-staining Visual Results

Figure 3(a)-(i) display the H&E-IHC stained image pairs with pseudo labels as ‘negative’, along with the generated IHC virtual-stained images (‘Generated IHC’). Figure 3(j)-(r) show the image pairs with pseudo labels as ‘positive’ and the generated IHC virtual-stained images. The results verify that our multi-modal pre-training framework can well capture the IHC-related information using only H&E images as inputs and generate corresponding brown/white IHC-stained regions.

#### 3.2 Intermediate Results in Sampling Process

In Figure 4, we present the intermediate results of our proposed framework during the sampling process of re-staining task. Following [1], given an H&E-stained image, we first extract its latent feature  $z_T = z_0^{he}$ . Then, for each  $t$  in range  $T$  to 1, we calculate the denoised feature map  $z_{t-1}$  through:

$$z_{t-1} = c_{zt}z_t + c_{yt}z_T - c_{\mu t}\mu_\theta(z_t, t) + \sqrt{\delta_t}\epsilon_t, \quad (1)$$

where  $c_{zt}$ ,  $c_{yt}$  and  $c_{\mu t}$  are computed as:

$$c_{zt} = \frac{\delta_{t-1}}{\delta_t} \frac{1 - m_t}{1 - m_{t-1}} + \frac{\delta_{t|t-1}}{\delta_t} (1 - m_{t-1}) \quad (2)$$

$$c_{yt} = m_{t-1} - m_t \frac{1 - m_t}{1 - m_{t-1}} \frac{\delta_{t-1}}{\delta_t} \quad (3)$$

$$c_{\mu t} = (1 - m_{t-1}) \frac{\delta_{t|t-1}}{\delta_t}, \quad (4)$$

where  $m_t = t/T$  and  $\delta_t$  is defined as  $2(m_t - m_t^2)$ .  $\epsilon_t \sim \mathcal{N}(\mathbf{0}, I)$  is Gaussian noise from a standard normal distribution. The overall timestep  $T$  in sampling process is set as 200.

We showcase some generated images at sampling steps 0, 40, 80, 120, 160 and 200, which correspond to denoised feature maps  $z_{200}, z_{160}, z_{120}, z_{80}, z_{40}$  and  $z_0$ , respectively. For example, the image generated after 40 sampling steps corresponds to  $z_{160}$ , the denoised feature map at the 160th time step. Figure 4(a)-(d) represent the intermediate results of generated images with the pseudo-labels as ‘positive.’ The results demonstrate that the our framework achieves high similarity with the ground truth in terms of color and structure after 160-200 sampling steps. Figure 4(e)-(h) show the intermediate results of generated images with the pseudo-labels as ‘negative.’ Similarly, after passing 160 sampling steps, the generated results become stable.

### REFERENCES

- [1] Bo Li, Kaitao Xue, Bin Liu, and Yu-Kun Lai. 2023. Bbmd: Image-to-image translation with brownian bridge diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*. 1952–1961.
- [2] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. 2021. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering* 5, 6 (2021), 555–570.



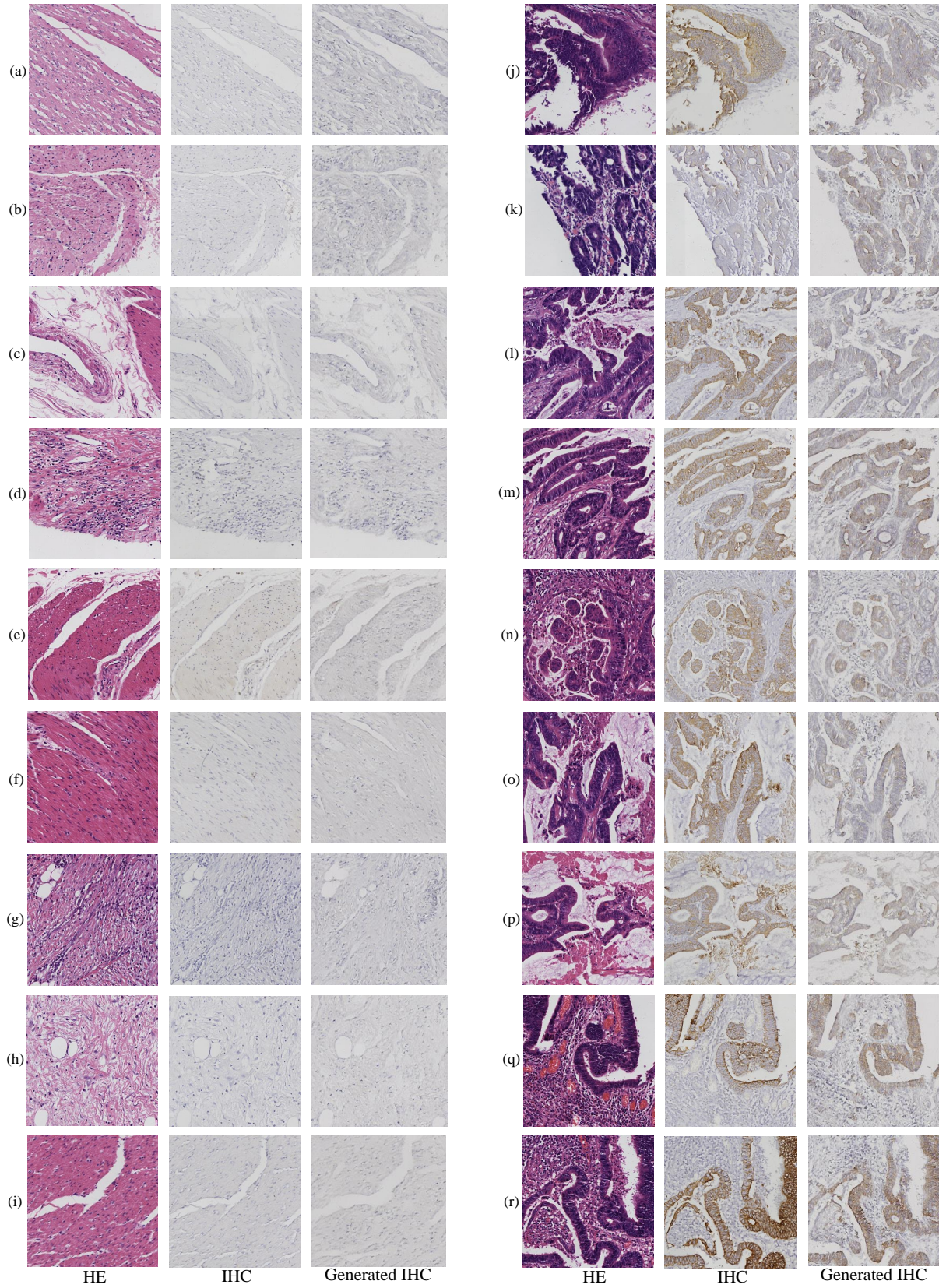
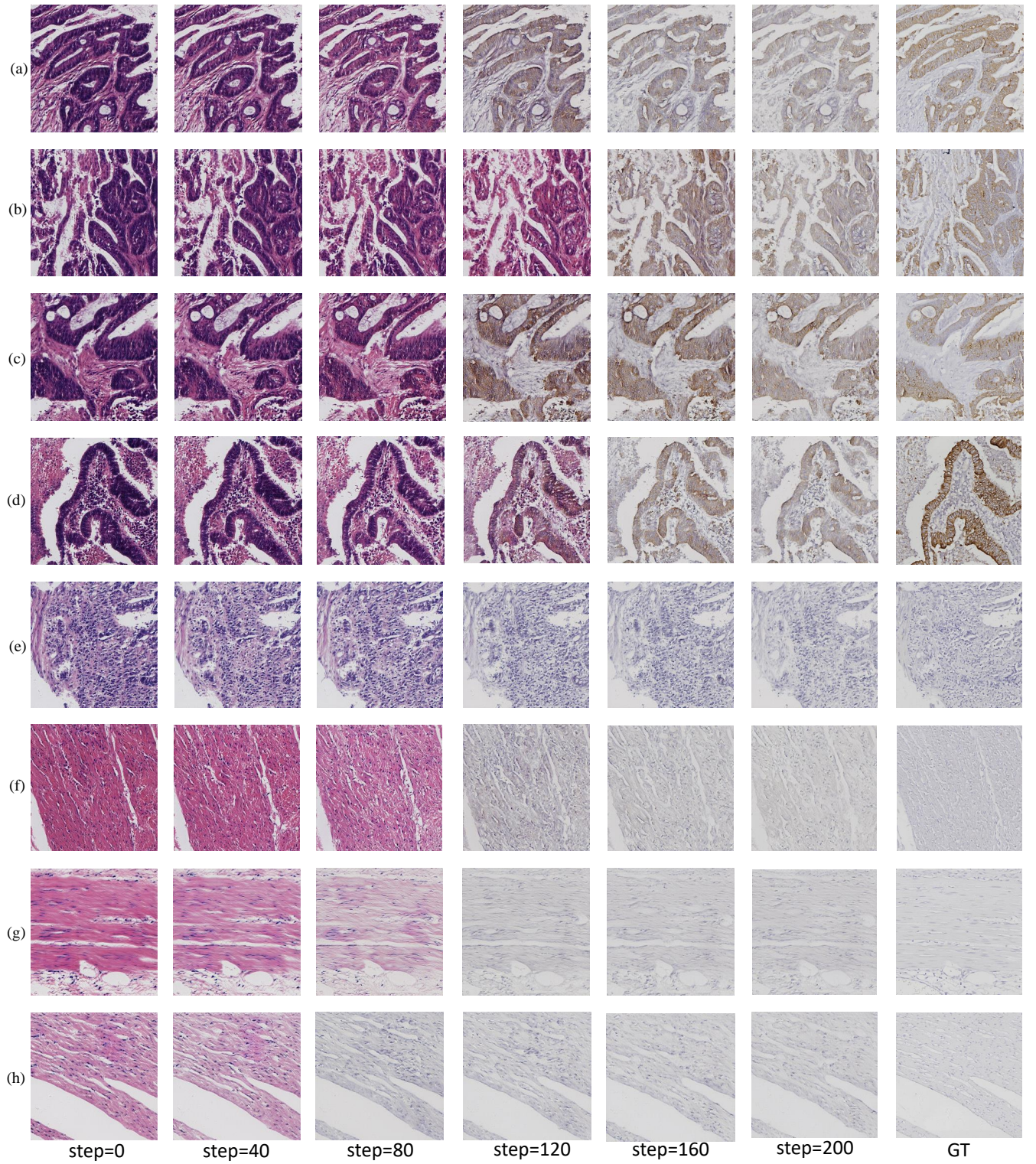


Figure 3: Visualization of the generated IHC virtual-stained image patches.





**Figure 4: The intermediate results of the sampling process of the re-staining task in generating the IHC virtual-stained image patches using HE-stained image patches as the inputs. The overall sampling steps is set as 200.**