Linearized MIL Attention Reveals a Traversable Direction in the Virchow2 Embedding Space

Ryoichi Koga[1 [0009−0008−7405−5553]✉], Tatsuya Yokota[1], Noriaki Hashimoto[2], Hiroaki Miyoshi[3], Ichiro Takeuchi[2,4], and Hidekata Hontani[1]

[1] Department of Computer Science, Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya-shi, Aichi 466-8555, Japan
koga@nitech.ac.jp
[2] RIKEN Center for Advanced Intelligence Project, 1-4-1 Nihonbashi, Chuo-ku, Tokyo, 103-0027, Japan
[3] Department of Pathology, Kurume University School of Medicine, 67 Asahi-cho, Kurume-shi, Fukuoka, 830-0011, Japan
[4] Department of Mechanical Systems Engineering, Nagoya University, Furo-cho, Chikusa-ku, Nagoya-shi, Aichi, 464-8601, Japan

Abstract. We propose a framework that explicitly links attention in multiple instance learning (MIL) to a specific direction vector in the Virchow2 embedding space and trains an embedding-conditioned image generator to render traversals along this direction. Many conventional MIL pipelines use ImageNet-pretrained models (e.g., ResNet) for embeddings, and their attention is computed as the inner product between a weight vector and an embedding after a non-linear transformation. In contrast, our proposed framework extracts patch embeddings using Virchow2, a pathology foundation model, and then computes the inner product between the embedding vector $\boldsymbol{h}$ and the weight vector $\boldsymbol{w}$ to determine attention strength. Since our framework does not involve non-linear transformations for embedding vectors, the learned direction $\boldsymbol{w}$ has a clear meaning as an attention-increasing direction in the Virchow2 embedding space. As a result, a linear traversal that increases attention can be represented in the embedding space for each patch. Our proposed framework trains a diffusion autoencoder conditioned on embeddings to render patch sequences along the linear traversal. These sequences provide a morphology-oriented visualization of changes associated with increasing attention.

Keywords: Attention Geometry · Foundation Model · Multiple Instance Learning · Diffusion Autoencoder

1 Introduction

Whole slide images (WSIs) provide morphological evidence of cellular tissue for cancer diagnosis [2]. WSIs are gigapixel-sized images, making accurate and exhaustive annotations of lesion regions impractical. Class labels used to construct a classifier are often available only at the slide level. To train classifiers under

such situations, multiple instance learning (MIL) [11] is widely adopted. In cancer diagnosis, explicitly stating the diagnostic basis through pathology-relevant factors, such as nuclear morphology and cell-type composition, is crucial.

Attention-based MIL (ABMIL) [8, 11, 12, 19, 20, 26, 27] provides heatmaps that highlight regions contributing to classification while avoiding annotations of lesion regions, making it one of the standard methods for WSI analysis. Here, these heatmaps only indicate which image patches are focused on for classification, not why. In other words, heatmaps show where the model focuses, yet do not reveal which morphological factors contribute to higher attention scores.

In many conventional ABMIL models, patch embedding vectors are non-linearly transformed to determine attention strength. This non-linear transformation makes it non-trivial to interpret the factors contributing to the attention. Many prior works [8,11,12,20,26,27] use ImageNet-pretrained models (e.g., ResNet) for patch embeddings, and the non-linear transformation for embedding vectors yields attention scores suitable for classification.

The emergence of pathology foundation models [4, 15, 24, 28] is transforming how algorithms for computer-aided pathological diagnosis are designed. Embeddings from foundation models for general images, such as natural images, support strong performance with linear readouts [5, 9, 17]. Similarly, pathology foundation models have begun to support high-accuracy linear readouts for (i) lesion detection and (ii) subtype classification based on lesion regions [4, 15, 24, 28]. This motivates us to explore whether pathology-relevant factors can be linearly accessed in such embeddings. Unlike many methods, which extract patch embeddings using an encoder trained on general images and then determine attention useful for pathological diagnosis via a non-linear transformation, our proposed framework extracts patch embeddings using Virchow2 [28], and then determines attention via a single linear operation. This allows the learned weight vector to be interpreted as a geometric direction within the embedding space.

In this paper, we propose a framework that explicitly links MIL attention to a specific direction vector in the Virchow2 embedding space and renders traversals along this direction with an embedding-conditioned image generator (Fig. 1). We call our framework, which determines attention strength via the inner product of the weight vector and the embedding vector, *Linearized Attention MIL (LAMIL)*. Since this framework does not involve non-linear transformations for embedding vectors, the learned weight vector explicitly indicates the direction that increases attention in the Virchow2 embedding space. In this embedding space, the linear traversal of each patch embedding vector along the weight vector yields a sequence with increasing attention scores. By rendering this sequence as pathological image patches with a diffusion autoencoder (DiffAE) and then applying nuclei segmentation and cell-type classification, we characterize changes in nuclear morphology and cell-type composition along the traversal. Our experimental results establish an explicit link between MIL attention and an attention-increasing direction in the Virchow2 embedding space. Our main contribution is an explicit and traversable geometric formulation of MIL attention in the Virchow2 embedding space.

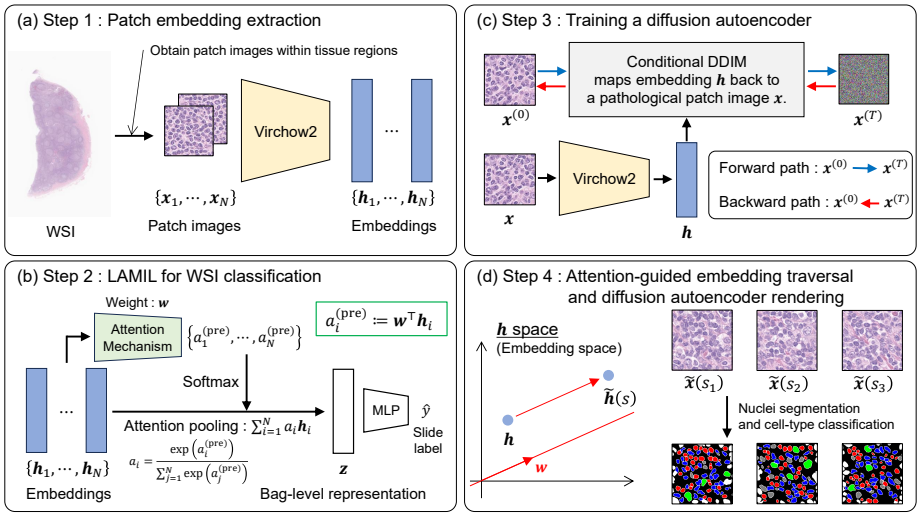

Fig. 1. Overview of the proposed framework. (a) Virchow2 produces patch embeddings from tissue regions within WSIs. (b) Linearized Attention MIL (LAMIL) performs slide-level prediction from the bag of embeddings. (c) A diffusion autoencoder (DiffAE) conditioned on embeddings is trained to reconstruct patch images. (d) At inference, LAMIL presents an attention-guided embedding traversal; edited embeddings are decoded by DiffAE to visualize attention-controlled patch sequences. These sequences can be further analyzed with nuclei segmentation and cell-type classification, providing a morphology-oriented visualization of changes associated with increasing attention in the rendered sequence.

2 Related Works

MIL has become a standard approach for weakly supervised WSI classification without requiring lesion annotations, and many works [8, 11, 12, 20, 26, 27] have improved performance by increasing the capacity of attention or aggregation mechanisms. Linear-attention-based MIL (Lin-MIL) has been explored to reduce the computational complexity of transformer self-attention [19]. In contrast, our goal is neither to increase attention capacity nor to approximate transformer self-attention. We apply a single linear scoring function directly to Virchow2 embeddings, without an intermediate non-linear attention network, such that the learned weight vector defines an explicit global editing direction.

Many conventional MIL methods use ImageNet-pretrained models to extract patch features [8, 11, 12, 20, 26, 27]. Pathology foundation models [4, 15, 24, 28] have begun to provide patch embeddings that support high performance even with simple readouts for pathological diagnosis. While sophisticated attention mechanisms remain valuable for performance, linear operations can become competitive and easier to interpret. Our framework leverages Virchow2 embeddings

to link MIL attention to a specific direction vector in the embedding space, thereby enabling explicit traversal of embeddings.

Generative models are extensively explored for controllable generation by mapping latent representations back to image space. A DiffAE achieves high-fidelity reconstruction conditioned on representations [18], and some diffusion-based methods [6, 25] demonstrate that pathological image patches can be reconstructed from patch representations. We use a DiffAE as a renderer, using conditional DDIM sampling to reconstruct pathological patches from embeddings. This renderer is essential for visualizing patch sequences with increasing attention scores.

In computational pathology, presenting explanations grounded in morphological factors of cellular tissue is preferable to region-level saliency alone [7,21]. Nuclei segmentation and cell-type classification enable nuclear-level quantification [10], achieving interpretation based on nuclear morphology and cell-type composition [13,16]. We analyze diffusion-rendered patch sequences along attention-guided traversals to characterize changes in cell-type composition, complementing region-level heatmaps with a morphology-oriented analysis.

3 Method

3.1 Multiple Instance Learning

Given a WSI, we obtain a set of N tissue patches $\{\boldsymbol{x}_i\}_{i=1}^N$ and then extract patch embeddings $\{\boldsymbol{h}_i\}_{i=1}^N$, where $\boldsymbol{x}_i \in \mathbb{R}^{W\times H\times C}$ is an input patch and $\boldsymbol{h}_i \in \mathbb{R}^d$ is a patch embedding. MIL models with an attention mechanism assign a normalized attention score $a_i \in \mathbb{R}$ to each instance and aggregate patch embeddings into a bag-level representation $\boldsymbol{z} = \sum_{i=1}^N a_i \boldsymbol{h}_i$, which is fed to a classifier to predict a slide label, where the normalized score a_i is obtained using a softmax function such that $\sum_{i=1}^N a_i = 1$. In gated-attention ABMIL [11], the logits (i.e., pre-softmax attention scores) are produced by a non-linear attention network:

$$a_i^{(\mathrm{pre})} = \boldsymbol{w}^\top \left(\tanh(\boldsymbol{V}\boldsymbol{h}_i) \odot \sigma(\boldsymbol{U}\boldsymbol{h}_i)\right), \tag{1}$$

where $\boldsymbol{w}$, $\boldsymbol{V}$ and $\boldsymbol{U}$ are learnable parameters, $\odot$ denotes element-wise multiplication, and $\sigma(\cdot)$ is the sigmoid function. Many subsequent MIL methods employ other non-linear transformations or higher-capacity aggregation mechanisms [8,12,20,26,27]. In our proposed framework, we instead use a single linear operation to obtain an explicit editing direction in the embedding space.

3.2 Diffusion Autoencoder

We employ a diffusion autoencoder (DiffAE) [18] as an embedding-conditioned renderer that maps an embedding $\boldsymbol{h}$ back to an image $\boldsymbol{x}$ (Fig. 1(c)), enabling visualization of traversal in the embedding space. Given a paired dataset of patch images and embeddings $\{(\boldsymbol{x}, \boldsymbol{h})\}$, DiffAE learns to reconstruct $\boldsymbol{x}$ conditioned on $\boldsymbol{h}$ by denoising from Gaussian noise.

Diffusion models are composed of a forward process and a backward process. The forward process can be described as the gradual addition of a small amount of Gaussian noise to input images over T steps, which results in a series of noisy images $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(T)}$ [18]. Given a time t, the forward process is defined as $q(\boldsymbol{x}^{(t)} \mid \boldsymbol{x}^{(t-1)}) = \mathcal{N}\Big(\sqrt{1-\beta^{(t)}}\,\boldsymbol{x}^{(t-1)},\, \beta^{(t)}\boldsymbol{I}\Big)$, where $\beta^{(t)}$ specifies the noise variance schedule. The corresponding noisy image of $\boldsymbol{x}^{(0)}$ at time t can be sampled from a distribution $q(\boldsymbol{x}^{(t)} \mid \boldsymbol{x}^{(0)}) = \mathcal{N}\Big(\sqrt{\alpha^{(t)}}\,\boldsymbol{x}^{(0)},\, \big(1-\alpha^{(t)}\big)\,\boldsymbol{I}\Big)$, where $\alpha^{(t)} = \prod_{s=1}^{t}\big(1-\beta^{(s)}\big)$. Diffusion models are constructed by learning the backward process, namely the distribution $p_\theta(\boldsymbol{x}^{(t-1)} \mid \boldsymbol{x}^{(t)})$, where θ are neural network parameters [18]. In denoising diffusion implicit models (DDIMs) [22], the forward process is defined as non-Markovian, and the computation for the corresponding backward process is given as the following update rule:

$$\boldsymbol{x}^{(t-1)} = \sqrt{\alpha^{(t-1)}}\Big(\frac{\boldsymbol{x}^{(t)} - \sqrt{1-\alpha^{(t)}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}^{(t)}, t)}{\sqrt{\alpha^{(t)}}}\Big) + \sqrt{1-\alpha^{(t-1)}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}^{(t)}, t), \quad (2)$$

where $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}^{(t)}, t)$ is the unconditional noise prediction network.

A DiffAE is constructed as a conditional DDIM decoder $\boldsymbol{\epsilon}_\theta\left(\boldsymbol{x}^{(t)}, t, \boldsymbol{h}\right)$ and performs the same DDIM update (Eq. (2)) with conditional noise prediction. This model is trained by solving the standard denoising objective [18, 25]:

$$\min_\theta \mathbb{E}_{t,\,\boldsymbol{x}^{(0)},\,\boldsymbol{\epsilon}}\left[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}^{(t)}, t, \boldsymbol{h})\right\|_2^2\right], \quad (3)$$

enabling reconstruction of $\boldsymbol{x}^{(0)}$ from noisy $\boldsymbol{x}^{(t)}$ guided by the embedding $\boldsymbol{h}$.

We use the DiffAE as a renderer that decodes an embedding into an image by sampling from the backward process conditioned on the embedding. In our framework, we apply DiffAE not only to the original embedding $\boldsymbol{h}$ but also to embeddings $\tilde{\boldsymbol{h}}$ obtained by attention-guided traversal (Fig. 1(d)), yielding pathological patches $\tilde{\boldsymbol{x}}$. These patches form rendered sequences whose appearance changes consistently with the intended embedding edits.

3.3 Proposed Framework

Virchow2 Embeddings. We employ Virchow2 [28] as a frozen embedding extractor, obtaining patch embeddings $\boldsymbol{h}_i \in \mathbb{R}^{2560}$ for all tissue patches $\boldsymbol{x}_i \in \mathbb{R}^{224\times224\times3}$ (Fig. 1(a)). These embeddings are expected to capture diverse and transferable pathological factors, providing a strong basis for MIL attention even with linear readouts.

LAMIL: Linearized Attention MIL. We introduce *linearized attention MIL (LAMIL)*, where the pre-softmax attention score for the i-th instance is computed by a single linear projection (Fig. 1(b)):

$$a_i^{(\mathrm{pre})} = \boldsymbol{w}^\top \boldsymbol{h}_i. \quad (4)$$

In ABMIL [11], attention is computed after a non-linear transformation of each patch embedding. In contrast, our framework uses patch embeddings $\boldsymbol{h}$ directly as *keys* in a dot-product attention with a single global *query* vector $\boldsymbol{w}$, making the direction of increasing attention explicit and input-independent.

We emphasize that this design is not motivated by simplicity alone. Compared to non-linear attention networks, linearization makes increasing (or decreasing) the attention score directly interpretable as a geometric direction within the embedding space, which is essential for defining an explicit and reproducible editing procedure when combined with an embedding-conditioned diffusion renderer described in Section 3.2.

Explicit Embedding Editing Rule and Attention-guided Traversal. Let $s \in \mathbb{R}$ denote an editing step size. Given a patch embedding $\boldsymbol{h}$ (e.g., selected by its attention quantile), we define an edited embedding by traversing along the learned weight vector $\boldsymbol{w}$:

$$\tilde{\boldsymbol{h}}(s) = \boldsymbol{h} + s\,\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}. \tag{5}$$

Since the pre-softmax attention score in Eq. (4) is linear in $\boldsymbol{h}$, positive and negative values of s in Eq. (5) increase and decrease the score, respectively, enabling a controlled traversal along the learned direction. In practice, we sample a set of steps $\{s_k\}_{k=1}^{K}$ to obtain a sequence of edited embeddings $\{\tilde{\boldsymbol{h}}(s_k)\}_{k=1}^{K}$.

Rendering and Nuclei-level Analysis. We render each edited embedding $\tilde{\boldsymbol{h}}(s_k)$ into a pathological patch using the DiffAE to obtain an attention-controlled patch sequence $\{\tilde{\boldsymbol{x}}(s_k)\}_{k=1}^{K}$. We further apply nuclei segmentation and cell-type classification to each rendered patch and characterize changes in nuclear morphology and cell-type composition using existing pipelines [13].

4 Experiments and Results

4.1 Experimental Setup

Datasets. We evaluate our approach primarily on a malignant lymphoma dataset consisting of 2,884 WSIs scanned at $\times 40$ magnification. This dataset includes three diagnostic subtypes: reactive lymphoid hyperplasia (Reactive: 752 cases), follicular lymphoma (FL: 1,009 cases), and diffuse large B-cell lymphoma (DLBCL: 1,123 cases). Subtype labels for all cases are provided by pathologists and are used as slide-level supervision for MIL training and evaluation.

To compare LAMIL and existing MIL methods on public WSI benchmarks, we use Camelyon16 [3] and TCGA-NSCLC [1]. Camelyon16 contains 399 sentinel lymph node WSIs with slide-level metastasis labels. TCGA-NSCLC comprises two TCGA projects (LUAD and LUSC) and includes 993 diagnostic WSIs (507 LUAD and 486 LUSC), as commonly reported in MIL benchmarks. On these public datasets, MIL is conducted as a binary classification problem. We follow the preprocessing protocol of TransMIL [20], performing tissue filtering and non-overlapping patch extraction at $\times 20$ magnification.

Table 1. MIL classification viability of linearized attention on public WSI benchmarks and the malignant lymphoma dataset. All methods use Virchow2 patch embeddings as a common backbone and are evaluated with 4-fold cross-validation. Their performance is indicated in the form of $(\text{mean})_{(\text{standard deviation})}$; the comparison focuses on aggregation/attention mechanisms rather than backbone quality or state-of-the-art classification performance.

	Camelyon16 (Public Dataset)		TCGA-NSCLC (Public Dataset)		Lymphoma (In-house Dataset)	
	Accuracy	F1	Accuracy	F1	Accuracy	Macro-F1
Mean-Pooling	$0.679_{(0.038)}$	$0.583_{(0.152)}$	$0.916_{(0.004)}$	$0.917_{(0.003)}$	$0.888_{(0.007)}$	$0.891_{(0.007)}$
Max-Pooling	$0.669_{(0.050)}$	$0.569_{(0.104)}$	$0.898_{(0.003)}$	$0.899_{(0.004)}$	$0.881_{(0.007)}$	$0.883_{(0.007)}$
ABMIL [11]	$0.855_{(0.058)}$	$0.835_{(0.048)}$	$0.930_{(0.019)}$	$0.931_{(0.019)}$	$0.897_{(0.007)}$	$0.901_{(0.006)}$
TransMIL [20]	$0.881_{(0.045)}$	$0.869_{(0.037)}$	$0.931_{(0.017)}$	$0.933_{(0.015)}$	$0.900_{(0.002)}$	$0.904_{(0.002)}$
DTFD-MIL (MAS) [26]	$0.797_{(0.050)}$	$0.748_{(0.033)}$	$0.931_{(0.017)}$	$0.932_{(0.017)}$	$0.899_{(0.003)}$	$0.902_{(0.003)}$
DTFD-MIL (AFS) [26]	$0.796_{(0.034)}$	$0.732_{(0.017)}$	$0.929_{(0.020)}$	$0.930_{(0.019)}$	$0.899_{(0.004)}$	$0.902_{(0.003)}$
ACMIL [27]	$0.866_{(0.034)}$	$0.839_{(0.024)}$	$0.930_{(0.014)}$	$0.931_{(0.014)}$	$0.898_{(0.004)}$	$0.902_{(0.003)}$
SCMIL [12]	$0.870_{(0.026)}$	$0.848_{(0.013)}$	$0.927_{(0.019)}$	$0.928_{(0.019)}$	$0.897_{(0.007)}$	$0.901_{(0.006)}$
LAMIL (Ours)	$0.881_{(0.028)}$	$0.866_{(0.018)}$	$0.932_{(0.012)}$	$0.933_{(0.011)}$	$0.900_{(0.006)}$	$0.903_{(0.005)}$

Implementation Details. For all datasets and all MIL methods, we use Virchow2 patch embeddings as a common backbone to ensure a fair comparison of aggregation procedures. We perform 4-fold cross-validation and report the mean evaluation scores across folds. All MIL models are trained using AdamW [14] with a learning rate of 1.0×10^{-4}, a batch size of 128, and 20 training epochs. These models are evaluated separately within each dataset.

To train a DiffAE, we use 288,000 paired samples of patch images and their Virchow2 embeddings. The DiffAE is trained using AdamW with a learning rate 1.0×10^{-4}, a batch size of 40, and 100 training epochs.

4.2 Classification Viability of Linearized Attention.

We assess whether replacing non-linear attention with a single linear projection causes a substantial degradation in WSI classification performance. The goal is not to improve classification accuracy, but to verify whether Virchow2 embeddings support an explicit global attention direction while maintaining performance comparable to established MIL aggregators. Table 1 demonstrates that LAMIL maintains performance comparable to modern attention-based MIL baselines despite using a single linear projection for the attention.

We hypothesize that this competitiveness is driven by the expressiveness of Virchow2 embeddings, which capture diverse and transferable pathological factors in a high-dimensional space and can be accessed via simple linear readouts. Importantly, this should not be taken to imply that sophisticated attention mechanisms are unnecessary; rather, it suggests that the design trade-off shifts under

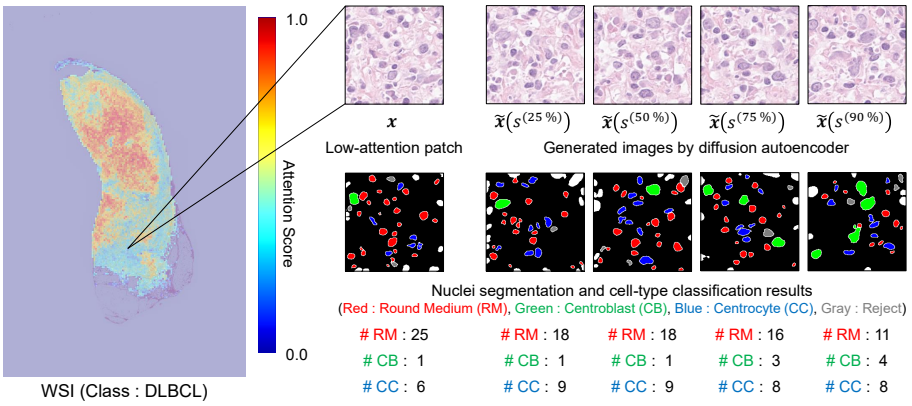

Fig. 2. Attention-guided embedding traversal and cell-type-level interpretation. Left: a WSI with the LAMIL attention heatmap using min-max normalized pre-softmax attention scores. Top-right: an original patch $\boldsymbol{x}$ selected from a low-attention quantile and diffusion-rendered patches conditioned on embeddings edited to match four target attention score quantiles. Bottom: nuclei segmentation and cell-type classification masks for the original and rendered patches.

transferable representations of Virchow2, enabling a controllable editing direction while maintaining classification viability. Whether other pathology foundation models support similarly usable global attention directions will be studied in future work.

4.3 Attention-guided Embedding Traversal and Patch Sequence Rendering

Fig. 2 illustrates a representative example of the proposed traversal. Starting from an original patch embedding $\boldsymbol{h}$ selected from a low-attention quantile, we edit $\boldsymbol{h}$ along Eq. (5) and decode the edited embeddings $\tilde{\boldsymbol{h}}(s)$ with DiffAE. Here, the initial noise for DiffAE is the same in the sequence. We set target attention levels to match the edited embeddings to prescribed attention score quantiles, yielding an ordered sequence whose attention increases across the selected quantiles. The rendered patches exhibit morphological transitions across the sequence, providing a concrete visualization of what increasing attention corresponds to in pixel space. The analysis in Fig. 2 is limited to a single representative DLBCL example and the positive attention direction. Therefore, it should be noted that it does not establish that the observed morphological trend generalizes across slides, diagnostic classes, cross-validation folds, or both traversal directions.

To interpret these transitions in terms of pathology, we apply nuclei segmentation and cell-type classification [13] to each rendered patch and visualize the resulting nuclei masks. DLBCL is characterized by the proliferation of large B-cells [23]. Across the rendered sequence, increasing attention is accompanied by an increased presence of centroblasts (CB: $1 \rightarrow 4$) and a decreased presence of

round medium cells (RM: 25 → 11), consistent with the large-cell morphology characteristic of DLBCL. This result provides a morphology-oriented characterization of a DiffAE-rendered attention traversal. Importantly, w represents a global attention direction shared across classes rather than a class-specific discriminative direction. Therefore, increasing the pre-softmax attention score along w does not necessarily imply an increase in the prediction score for DLBCL or any other specific class. The example in Fig. 2 should therefore be interpreted as a morphology-oriented characterization of increasing LAMIL attention in a DLBCL slide, rather than as a traversal toward the DLBCL class. It should be noted that verifying whether the generated patches faithfully reproduce real tissue transformations remains an open question.

5 Conclusion

We present a framework that explicitly links MIL attention to a direction vector in the embedding space and trains an embedding-conditioned image generator, thereby providing a morphology-oriented visualization of changes along the learned attention direction. Using Virchow2 embeddings, LAMIL provides a straightforward editing rule for traversing the learned attention direction, while DiffAE rendering and nuclei-level analysis characterize the resulting sequence in terms of nuclear morphology and cell-type composition. Our framework provides an explicit and traversable geometric formulation of MIL attention in the Virchow2 embedding space.

Acknowledgments. This work was supported by the Kayamori Foundation of Informational Science Advancement and the Naito Research Grant to R.K., and by JSPS KAKENHI Grant Numbers 25K24409 and 26K21248 to R.K. and 25K03139 to H.H.

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

References

1. The cancer genome atlas pan-cancer analysis project. Nature Genetics, **45**(10), 1113–1120, (2013)
2. National cancer institute: What is cancer? (2021), https://www.cancer.gov/about-cancer/understanding/what-is-cancer, (Last Accessed on 20 Feb. 2026)
3. Bejnordi, B., Veta, M., Johannes, v.D., et al.: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. The Journal of the American Medical Association. **318**(22), 2199–2210, (2017)
4. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Chen, B., Zhang, A., Shao, D., Song, A.H., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature Medicine, **30**, 850–862, (2024)
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning. vol. 119, pp. 1597–1607, (2020)

6. Graikos, A., Yellapragada, S., Le, M.Q., Kapse, S., Prasanna, P., Saltz, J., Samaras, D.: Learned representation-guided diffusion models for large-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8532–8542, (2024)
7. Graziani, M., Andrearczyk, V., Marchand-Maillet, S., Müller, H.: Concept attribution: Explaining cnn decisions to physicians. Computers in Biology and Medicine, **123**, 103865, (2020)
8. Hashimoto, N., Fukushima, D., Koga, R., Takagi, Y., Ko, K., Kohno, K., Nakaguro, M., Nakamura, S., Hontani, H., Takeuchi, I.: Multi-scale domain-adversarial multiple-instance cnn for cancer subtype classification with unannotated histopathological images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3852–3861, (2020)
9. He, K., Chen, X., Xie, S., Li, Y., Doll'ar, P., Girshick, R.B.: Masked autoencoders are scalable vision learners. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15979–15988, (2022)
10. Hörst, F., Rempe, M., Heine, L., Seibold, C., Keyl, J., Baldini, G., Ugurel, S., Siveke, J., Grünwald, B., Egger, J., Kleesiek, J.: Cellvit: Vision transformers for precise cell segmentation and classification. Medical Image Analysis, **94**, 103143, (2024)
11. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: Proceedings of the 35th International Conference on Machine Learning. vol. 80, pp. 2127–2136, (2018)
12. Juyal, D., Shingi, S., Javed, S., Padigela, H., Shah, C., Sampat, A., Khosla, A., Abel, J., Taylor-Weiner, A.: Sc-mil: Supervised contrastive multiple instance learning for imbalanced classification in pathology. In: IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 7931–7940, (2024)
13. Koga, R., Koide, S., Tanaka, H., Taguchi, K., Kugler, M., Yokota, T., Ohshima, K., Miyoshi, H., Nagaishi, M., Hashimoto, N., Takeuchi, I., Hontani, H.: A study of criteria for grading follicular lymphoma using a cell type classifier from pathology images based on complementary-label learning. Micron, **184**, 103663, (2024)
14. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: Proceedings of International Conference on Learning Representations, (2019)
15. 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**, 863–874, (2024)
16. Nagaishi, M., Miyoshi, H., Kugler, M., Sato, K., Kohno, K., Takeuchi, M., Yamada, K., Furuta, T., Hashimoto, N., Takeuchi, I., Hontani, H., Ohshima, K.: The detection of neoplastic cells using objective cytomorphologic parameters in malignant lymphoma. Laboratory Investigation, **104**(3), 100302, (2024)
17. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision. Transactions on Machine Learning Research, (2024)
18. Preechakul, K., Chatthee, N., Wizadwongsa, S., Suwajanakorn, S.: Diffusion autoencoders: Toward a meaningful and decodable representation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10619–10629, (2022)
19. Richter, C., Reisenbüchler, D., Schaadt, N.S., Feuerhake, F., Merhof, D.: Linear attention-based multiple instance learning for computational pathology. In: Proceedings of the MICCAI Workshop on Computational Pathology. pp. 86–96, (2025)

20. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., Zhang, Y.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. In: Proceedings of Advances in Neural Information Processing Systems (2021)
21. Shuting, X., Junlin, H., Hao, C.: Explain any pathological concept: Discovering hierarchical explanations for pathology foundation models. In: Proceedings of Medical Image Computing and Computer Assisted Intervention, (2025)
22. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: Proceedings of International Conference on Learning Representations, (2021)
23. Swerdlow, S.H., Campo, E., Harris, N.L., Jaffe, E.S., Pileri, S.A., Stein, H., Thiele, J.: World Health Organization classification of tumours of haematopoietic and lymphoid tissues. Revised 4th ed. International Agency for Research on Cancer, (2017)
24. Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., Xu, Y., Wei, M., Wang, W., Ma, S., Wei, F., Yang, J., Li, C., Gao, J., Rosemon, J., Bower, T., Lee, S., Weerasinghe, R., Wright, B.J., Robicsek, A., Piening, B., Bifulco, C., Wang, S., Poon, H.: A whole-slide foundation model for digital pathology from real-world data. Nature, **630**, 181–188, (2024)
25. Xu, X., Kapse, S., Gupta, R., Prasanna, P.: Vit-dae: Transformer-driven diffusion autoencoder for histopathology image analysis. In: Deep Generative Models: Third MICCAI Workshop, DGM4MICCAI 2023, Held in Conjunction with MICCAI 2023. pp. 66–76, (2023)
26. Zhang, H., Meng, Y., Zhao, Y., Qiao, Y., Yang, X., Coupland, S.E., Zheng, Y.: Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18780–18790, (2022)
27. Zhang, Y., Li, H., Sun, Y., Zheng, S., Zhu, C., Yang, L.: Attention-challenging multiple instance learning for whole slide image classification. In: European Conference on Computer Vision. pp. 125–143, (2024)
28. Zimmermann, E., Vorontsov, E., Viret, J., Casson, A., Zelechowski, M., Shaikovski, G., Tenenholtz, N., Hall, J., Fuchs, T., Fusi, N., Liu, S., Severson, K.: Virchow2: Scaling self-supervised mixed magnification models in pathology. arXiv:2408.00738, (2024)