SlideMix: Enhancing Whole Slide Image Analysis via Multimodal Shuffling

Anonymous CVPR submission

Paper ID 11782

Abstract

Pathological whole slide images (WSIs) are pivotal in cancer diagnosis, but their gigapixel scales and complex heterogeneity make the manual review slow and may lead to inconsistent conclusions. Deep learning studies are therefore applied yet three obstacles persist in WSI analysis: (i) weak supervision from slide-level labels without region-level guidance, (ii) locally homogeneous yet globally heterogeneous tissue that complicates spatial reasoning, and (iii) cross-magnification patterns that call for effective multi-scale pattern modeling. Existing architectural innovations' improvements remain limited and inconsistent across the benchmarking datasets. The recent augmentations (e.g., CutMix) draw new insight but only partially address these issues and introduce label-irrelevant supervision signals. We propose SlideMix, a model-agnostic multimodal augmentation framework for WSI backbones that: (1) employs a VLM-based Visual-language Adaptive Region (VAR) selector to mitigate weak-label noise by prioritizing diagnostically relevant regions; (2) performs In-place Tile Shuffling (ITS) to balance local homogeneity with global heterogeneity without breaking slide context; and (3) integrates a multi-factor, loss-driven, online Curriculum-Learning Feedback (CLF) scheme for progressive cross-scale representation learning. Across 11 WSI benchmarks, SlideMix consistently improves accuracy and generalization over strong state-of-the-art backbones. The extensive experiments highlight SlideMix as a simple, plug-and-play route to more robust and scalable digital pathology models. The project will be open-sourced.

1. Introduction

Pathology diagnosis serves as the gold standard for cancer diagnosis, relying on microscopic image interpretation to ensure accurate disease identification and treatment planning [14]. In modern digital pathology, tissue biopsies are routinely digitized into gigapixel-scale Whole Slide Images (WSIs), which preserve rich spatial details and multi-scale tissue features. However, their massive size makes manual review both labor-intensive and dependent on highly specialized expertise, often leading to inconsistencies and variability in diagnostic outcomes [18].

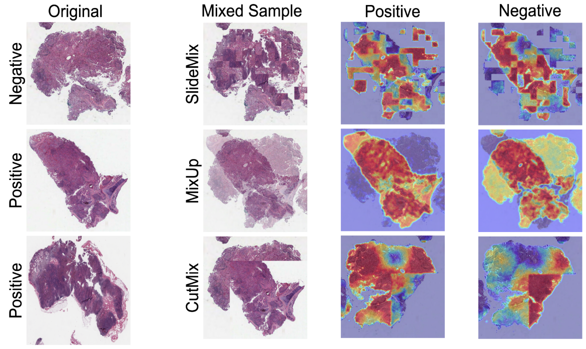

Figure 1. Visualization of an augmented example. Compared to other mixing-based methods, SlideMix accurately identifies tile and instance boundaries. It further distinguishes effectively between negative and positive samples.

Deep learning techniques have been increasingly adopted to assist in WSI analysis by automatically identifying and categorizing critical histological patterns [35, 41]. Recent methods currently employ Multiple Instance Learning (MIL), a two-stage framework designed to overcome GPU memory limits imposed by gigapixel-scale slides [31]. Specifically, in the first stage of MIL, WSIs are divided into smaller tiles (e.g., 224×224 pixels from a $100,000 \times 80,000$-pixel WSI), and encoded into feature embeddings using a tile-level foundation model (e.g. UNI [7]). In the second stage, the tile embeddings are aggregated into bags to produce slide-level predictions with a slide-level model (e.g. TransMIL [28]) [40, 42]. Although practical, the inherent complexity and multi-scale nature of WSIs continue to challenge existing models, particularly in their ability to effectively integrate cross-scale information [35]. Three core challenges can thus be identified:

1) ***Weak supervision from scarce fine-grained annotations*** [13]. Unlike natural images, only a small fraction of regions within a WSI (often $< 1\%$) are label-relevant. This extreme imbalance between informative and non-informative regions greatly complicates the learning

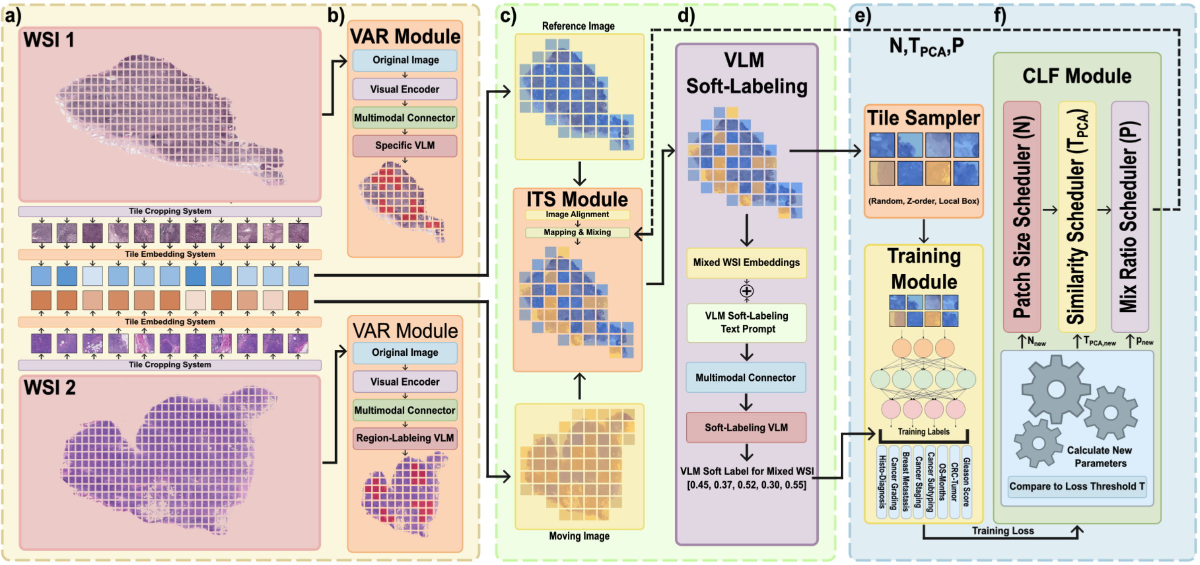

Figure 2. Overview of the MIL pipeline integrated with SlideMix. (a) Two WSIs from the same dataset are selected, tiled, and embedded into feature vectors for efficient computation. (b) The corresponding low-resolution WSIs are input into a fine-tuned VLM, which identifies label-relevant regions for mixing. (c) Coordinates of the selected candidate regions are passed to the data augmentation module, where the embedded tiles from stage one of MIL are shuffled in place according to three adjustable parameters N, T_{PCA}, and P provided by the CLF, creating a new mixed sample. (d) The mixed sample is then processed by another VLM to generate a soft mixed label. (e) This new sample–label pair is fed into the tile sampler and subsequent training module. (f) Finally, the training loss is used to update the three aforementioned CLF parameters, enabling the ITS to automatically adjust the difficulty of the mixed samples throughout training.

of discriminative features, as irrelevant tiles dominate the training process and weaken label-feature alignment.

2) ***Spatial modeling impaired by local homogeneity and global heterogeneity*** [42]. Pathological structures exhibit strong local continuity but substantial global diversity. This dual characteristic complicates spatial reasoning, as models must simultaneously preserve fine-grained local consistency while capturing long-range dependencies across the entire slide—a challenge that becomes increasingly pronounced with larger WSI dimensions.

3) ***Multi-scale nature against limited feature fusion capability*** [8]. WSIs contain features spanning multiple scales, from cellular to organ-level structures, and different diagnostic tasks emphasize distinct scales. For instance, tumor purity regression depends on global contextual understanding, while EGFR mutation prediction demands precise modeling of cellular morphology.

To address these challenges, prior studies have explored architectural innovations [8, 15] and data augmentation methods [11, 25, 29]. However, these approaches often overlook key properties of pathological images (Fig. 1), including their multi-scale nature and severe label imbalance caused by large proportions of irrelevant regions. Moreover, most augmentation techniques remain static, limiting adaptability across diverse WSIs and tasks.

To this end, we propose SlideMix, a novel multimodal data augmentation framework for MIL-based WSI analysis (Fig. 2). SlideMix integrates a Visual-Language Model (VLM)-guided region selector, an in-place tile shuffling mechanism, and an adaptive curriculum feedback loop. Together, these components address the aforementioned challenges through three key innovations:

1) We propose a **VLM-based Adaptive Region (VAR) Selector** that employs Retrieval-Augmented Generation (RAG) to retrieve domain-relevant knowledge and identify diagnostically significant ROIs at the lowest WSI scale, mitigating weak supervision effects.

2) We design an efficient **In-place Tile Shuffling (ITS) module** that mixes tile embeddings between WSIs using ROI coordinates from the VARS, balancing local homogeneity and global heterogeneity. A VLM generates soft labels from the mixed WSIs to create new training data.

3) We introduce a **multi-factor Curriculum Learning Feedback (CLF) module** that adaptively adjusts the shuffle ratio, PCA similarity threshold, and shuffle granularity in the ITS module based on loss evaluation, enabling progressive cross-scale feature learning.

Experimental results show that SlideMix improves performance over 8 pathological tasks across 10 state-of-the-art WSI models on 11 pathological WSI datasets with 20,523 WSIs, demonstrating its robustness and adaptability. By providing new insights into multimodal feature regrouping, SlideMix improves model generalization and diagnostic accuracy in digital pathology applications.

2. Related Works

2.1. Data Augmentation

Data augmentation is a cornerstone technique in deep learning, designed to enhance model generalization and robustness by artificially expanding the training dataset. For natural images, common methods include simple geometric and color transformations, such as flipping, rotating, and color jittering [6]. However, these generic augmentations often fail to capture the unique and complex characteristics of pathological images. Consequently, domain-specific augmentation techniques have been developed. Early pathology-specific methods operated at the tile-level, including stain normalization [11] and the generation of synthetic artifacts using GANs [27] or latent-space [24] models. Although more relevant to pathology, these methods overlook the broader spatial context and multi-scale features inherent in a WSI. More recent methods have shifted toward feature-level augmentation in the second stage of MIL to better model spatial relationships. For instance, Z-Order sampling [25] processes tiles in Z-order to preserve the smallest spatial distance between the tiles, encouraging the model to learn spatial dependencies. However, such methods remain label-agnostic, applying uniformly across the entire WSI, and inadvertently mixing large label-irrelevant regions with the few diagnostically critical ones. This introduces substantial noise and hinders the model's ability to learn discriminative features.

2.2. Visual Language Models

Visual-Language Models (VLMs) have demonstrated remarkable capabilities in bridging vision and natural language, enabling complex reasoning that requires multi-modal understanding[3, 39]. These models are often pre-trained on large datasets of image-text pairs and excel at zero-shot generalization for tasks such as image classification, object detection, and visual question answering. In the medical domain, VLMs are increasingly applied to interpret complex medical imagery by leveraging associated textual information, such as clinical notes or pathology reports [19, 32]. Their ability to ground textual concepts within visual data makes them particularly promising for localizing ROIs in WSIs. However, most pathology applications of VLMs have focused on direct diagnosis or report generation. In contrast, our approach uses a VLM not as a direct diagnostic tool, but as an intelligent guidance mechanism.

2.3. Curriculum Learning

Curriculum learning is a training strategy inspired by human cognition, in which models are presented with training examples in a structured order [2]. The model first learns simpler concepts and then progressively tackles more difficult ones, which facilitates faster convergence and improved performance. However, due to the complex and heterogeneous nature of WSIs, defining an effective curriculum in computational pathology is challenging. Previous WSI data augmentation methods applied fixed augmentation rules throughout training [25], failing to adapt to the model's evolving learning state or to the varying complexity of different WSIs and diagnostic tasks.

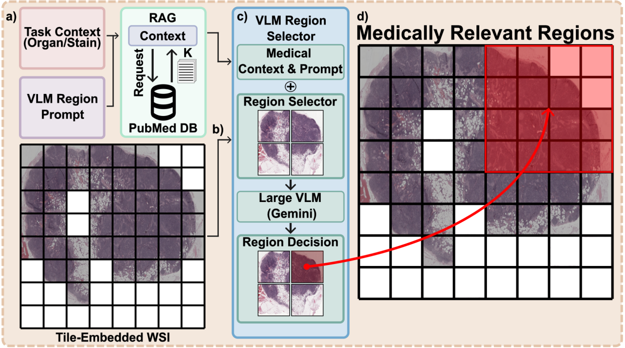

Figure 3. The proposed VARS: a) The task label L and a visual-language model (VLM) region prompt P are sent to a retrieval-augmented generation (RAG) system R to retrieve the top-K relevant medical data sources D from PubMed $\mathcal{D}$. b) The WSI S at lowest magnification is processed by the visual encoder G_v of the modified Gemini model Gem_f to extract visual features F, which are partitioned into large regions. c) F, L, and D are fused via the multimodal connector G_m within Gem_f. d) Gemini predicts outcomes for each region, while the coordinate generator G_p proposes coordinates C for potential regions of interest (ROIs).

3. Methods

3.1. Data Pre-processing and Tile Embedding

A WSI S is first loaded at a target microns-per-pixel (mpp) resolution, then partitioned into a non-overlapping grid of tiles $\{T_{\text{i,j}}\}$ of a corresponding size T_{size} (Fig. 2a). A two-stage filtering protocol is employed to ensure the quality of the selected regions. It first discards tiles with tissue coverage below a predefined threshold (e.g., $< 50\%$ of the tile area), then removes tiles where the pixel variance falls below a quantitative cutoff (e.g., $Var(I) < 0.01$, where I represents the pixel intensity normalized to the [0,1] range.). This ensures only valid tiles are retained for the downstream. Following pre-processing, each filtered tile is embedded into a dense, semantic feature vector (Fig. 2a) using a pathological foundation model (e.g., UNI [7]). This embedding process boosts both training efficiency and performance of the slide-level model (e.g., ABMIL [15]).

3.2. VLM-based Adaptive Region Selector

To dynamically select the label-related regions in WSIs, we propose VLM-based Adaptive Region Selector (VARS) (Fig. 3). Instead of conventional static selectors, VARS dynamically analyzes the WSIs, proposing the coordinates of

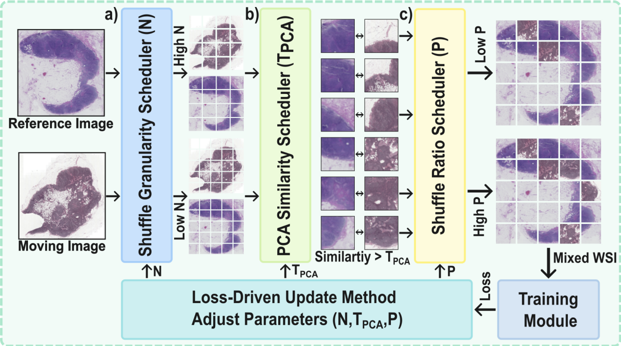
a)
b)
c)
Reference Image
Moving Image
Shuffle Granularity Scheduler (N)
High N
Low N
PCA Similarity Scheduler (TPCA)
Similartiy > TPCA
Shuffle Ratio Scheduler (P)
Low P
High P
Mixed WSI
↑N
↑TPCA
↑P
Loss
Loss-Driven Update Method
Adjust Parameters (N,TPCA,P)
Training
Module

Figure 4. The proposed CLF module dynamically adjusts three schedulers based on training loss l and threshold T_{loss}. a) Shuffle Granularity (N) controls shuffling region scale. b) PCA similarity threshold (T_{PCA}) governs feature similarity between tiles selected for shuffling. c) Shuffle Ratio (P) sets the proportion of shuffled tiles. These updated parameters are then passed to the ITS module to configure the next augmentation.

the regions related to the labels. These proposals significantly enhance the impact of the downstream shuffling process. In VARS, we integrate a VLM, Gemini [30], with a Retrieval-Augmented Generation (RAG) system.

In a single augmentation process, two VARS will handle two WSIs from the same dataset simultaneously (Fig. 2b). In a VARS, a WSI S is loaded at the lowest magnification level (highest mpp) for efficiency, then passed into the visual encoder $\mathrm{G_v}$ of the modified Gemini $\mathrm{Gem_f}$ to generate the visual features F:

$$F = \mathrm{G_v}(S) \tag{1}$$

Accordingly, the task labels L (e.g., OS Months: 36.5) of this WSI and a VLM prompt P (e.g., "Identify key pathological regions in this WSI.") are sent together into RAG R. It allows Gemini to retrieve knowledge from the PubMed database $\mathcal{D}$ [5] in real-time, enhancing its performance:

$$D = \mathrm{R}_{\mathcal{D}}(L, P) \tag{2}$$

After that, the multimodal connector $\mathrm{G_m}$ of $\mathrm{Gem_f}$ first fuses F, L, and D. This combined representation is then passed to the predictor $\mathrm{G_p}$ (also of $\mathrm{Gem_f}$) to propose the task-specific regions coordinates C:

$$C = \mathrm{G_p}(\mathrm{G_m}(F, L, D)) \tag{3}$$

The result C guides the downstream augmentation modules to apply shuffling only to the selected meaningful regions.

3.3. Curriculum Learning Feedback Module

To enable gradual cross-scale feature learning, we implement the Curriculum Learning Feedback (CLF) module (Fig. 4). It dynamically adjusts the difficulty of the data augmentation in the downstream ITS module based on the model's performance. In CLF, the curriculum starts with simple augmentations and progressively increases the difficulty by three schedulers controlling shuffle granularity, PCA similarity threshold, and shuffle ratio. The difficulty is set by comparing the model's current training loss l against a performance threshold T_{loss}.

Shuffle Granularity (N): This scheduler determines the feature scale the model learns to recognize. It first groups tile embeddings into different sizes, denoted as $N \times N$, The curriculum starts with large N (e.g., 16 × 16), a relatively easy task that lets the model learn the coarse-grained features (e.g., the boundaries between different tissues). It then gradually decreases N in the sequence of $[N_0, N_1, ..., N_n]$, where $N_{i+1} < N_i$. In the end, the model will learn from the most fine-grained patterns (e.g., the features inside a cell).

PCA Similarity Threshold (T_{PCA}): This scheduler determines the feature similarity the model learns to recognize. Tile embeddings close in PCA space are also close in label space [22]. We define the distance $PCA_{\Delta}(e_i, e_j)$ between the principal components of two tiles e_i and e_j as:

$$PCA_{\Delta}(e_i, e_j) = |\mathrm{PCA}(e_i) - \mathrm{PCA}(e_j)|^2 \tag{4}$$

To ensure a consistent difficulty scale, we min-max normalize this distance to $[0, 1]$ using the statistics (PCA_{min}, PCA_{max}) from the entire training set $\mathcal{D}$:

$$PCA'_{\Delta} = \frac{PCA_{\Delta} - PCA_{min}}{PCA_{max} - PCA_{min}} \tag{5}$$

A threshold $T_{PCA} \in [0, 1]$ constrains the shuffling: a tile pair can be shuffled only if $PCA'_{\Delta}(e_i, e_j) < T_{PCA}$. The curriculum starts with a relatively high T_{PCA} (allowing dissimilar tiles to be shuffled), then gradually decreases the PCA similarity threshold, $[T_{PCA,0}, ..., T_{PCA,n}]$, where $T_{PCA,i+1} < T_{PCA,i}$. At the end, the model is forced to learn fine-grained differences, as only the most similar tiles are permitted to be shuffled.

Shuffle Ratio (P): This scheduler determines the feature integrity the model learns to recognize. It controls the shuffle ratio, denoted as $P \in [0, 1]$. The curriculum starts with a low P, a relatively easy task that lets the model learn from relatively intact features (e.g., 90% integrity). It then gradually increases P in the sequence of $[P_0, P_1, ..., P_n]$, where $P_{i+1} > P_i$. In the end, the model will learn from the highly fragmented features.

The parameters of all three schedulers are updated at the end of each training epoch. We implement the loss-hold strategy as the update rule. Let $\boldsymbol{\Theta}_i = (P_i, N_i, T_{PCA,i})$ represent the curriculum parameters at epoch i. The parameters for the subsequent epoch $\boldsymbol{\Theta}_{i+1}$ are:

$$\boldsymbol{\Theta}_{i+1} = \begin{cases} (P_{i+1}, N_{i+1}, T_{PCA,i+1}) & \text{if } l < T_{loss} \\ \boldsymbol{\Theta}_i & \text{otherwise} \end{cases} \tag{6}$$

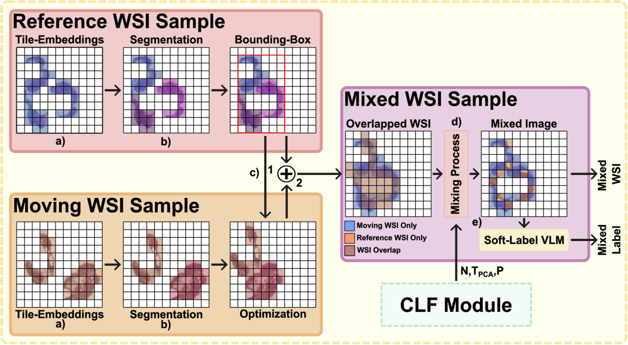

Figure 5. The ITS module operates as follows: a) a pair of embedded WSIs are provided and assigned as ‘reference’ and ‘moving’; b) each WSI is segmented into biological objects approximated by convex hulls; c) the reference WSI is enclosed within its minimal bounding box, and the moving WSI aligns to maximize overlap; d) overlapping regions are shuffled based on ITS configuration and CLF parameters; and e) soft labels for the mixed image are generated using a VLM, forming a new training sample.

This ensures difficulty only escalates after the model masters the current difficulty level.

3.4. In-place Tile Shuffling Module

To dynamically shuffle tiles and generate corresponding soft labels, we introduce the In-place Tile Shuffling (ITS) module (Fig. 5). It dynamically moves instances in a WSI pair to create the maximum overlaps for better shuffling, and generate soft-labels on the augmented WSI with a VLM to guarantee the accuracy of the labels.

The process starts with a input raw WSI pair from ITS, whose roles are determined as the static reference WSI (S_r) and the moving WSI (S_m).

To create the maximum overlaps for a meaningful WSI augmentation, we align the biological structures of S_m to S_r before shuffling. The spatial layout of a WSI S is represented as a set of tile coordinates $T \subset \mathbb{Z}^2$. T is partitioned into disjoint, connected components $\{z_1, z_2, ..., z_n\}$, where each z_i corresponds to a distinct biological structure (e.g., a contiguous tissue section):

$$T = \bigcup_{i=1}^{n} z_i, \quad \text{where } z_i \cap z_j = \emptyset \text{ for } i \neq j \tag{7}$$

Each component z_i is treated as an independently movable object. We seek an optimal set of 2D translation vectors $\Theta = \{\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_n\}$ for S_m's components ($z_{m,i}$) to maximize the spatial overlap with S_r's coordinates T_r:

$$\Theta^* = \arg\max_{\Theta} \left| \left(\bigcup_{i=1}^{n} (z_{m,i} + \boldsymbol{\theta}_i) \right) \cap T_r \right| \tag{8}$$

Directly optimizing this objective on millions of tiles is computationally prohibitive. To create a tractable problem, we approximate each tissue component z_i with its convex hull $\mathrm{Conv(z_i)}$. A standard optimizer then solves for the translations Θ^* that align the convex hulls. Once found, these optimal offsets are applied to the original tile coordinates of S_m to produce an aligned WSI S_a.

The shuffling process occurs within the aligned, overlapping region $T_r \cap T_m$ in S_a. The coordinates of tile embeddings from S_m and S_r are randomly shuffled due to the configurations (P, N, T_{PCA}) from ITS. This shuffled WSI S_s contains a semantically coherent fusion of tissue structures from both S_r and S_m.

We then integrate CONCH [21], a pathological VLM, with our RAG system as a soft-labeler to generate the corresponding soft labels L_s. Similar to the Gemini in VARS, CONCH sends the labels and prompts to the PubMed database and retrieves medical knowledge for assistance. The CONCH VLM uses a different embedding system, so the same mixed image is re-recalculated with different tile embeddings only for soft-labeling.

3.5. Slide-level Feature Modeling

Augmented WSI S_s and labels L_s are passed to the slide-level backbone (e.g., TransMIL [28]) to generate the final slide-level predictions for different downstream tasks, following the conventional MIL pipeline. At the end of the epoch, the validation loss is sent back to CLF to adjust the parameters for the next augmentation.

4. Experiment

4.1. Datasets and Downstream Tasks

To demonstrate its effectiveness and generalizability, we evaluated SlideMix on 11 datasets across eight downstream tasks covering diverse diagnostic scenarios (Fig. 6). The **PANDA** [4] dataset with the **Gleason Score** task assesses prostate cancer aggressiveness, while **CAMELYON16** [1] with the **Breast Metastasis** task classifies lymph nodes as normal or tumorous. **IMP-CRS-2024** [23] with the **CRC-Tumor** task identifies tumor tissues in colorectal images. The **TCGA** [10] datasets cover multiple cancer types and tasks: **TCGA-Lung** performs **Cancer Subtyping**, distinguishing lung cancer variants; **TCGA-BLCA** performs **Cancer Staging**, assessing bladder cancer progression; **TCGA-UCEC** and **TCGA-CESC** perform **Cancer Grading** on uterine and cervical tissues, respectively, determining tumor differentiation levels; **TCGA-UCS** and **TCGA-UVM** perform **Histological Diagnosis**, classifying uterine and uveal tissue subtypes based on morphology; and **TCGA-BRCA** and **TCGA-GBM** perform the **OS-Months** task, predicting patient survival time in months from breast and brain tissue morphology, respectively.

Table 1. Benchmarking SlideMix Against SOTA Augmentation Methods With ABMIL.

Method	PANDA	CAMELYON16	IMP-CRS-2024	TCGA-Lung	TCGA-BLCA	TCGA-UCEC	TCGA-CESC	TCGA-UCS	TCGA-UVM	TCGA-BRCA	TCGA-GBM
	Acc. [%]	Acc. [%]	Acc. [%]	Acc. [%]	Acc. [%]	Acc. [%]	Acc. [%]	Acc. [%]	Acc. [%]	Corr.	Corr.
Baseline	75.9	93.7	95.2	71.6	48.4	67.5	54.4	58.3	62.5	0.449	0.614
MixUp	76.3	94.1	95.4	72.1	49.2	68.1	54.8	59.1	63.2	0.461	0.625
CutMix	76.8	94.5	95.6	72.5	49.7	68.6	55.2	59.8	63.8	0.472	0.628
CutOut	75.2	93.1	94.8	71.2	47.8	66.9	53.9	57.5	61.8	0.438	0.608
ResizeMix	77.1	94.8	95.8	72.9	50.1	69.1	55.6	60.3	64.2	0.478	0.632
PuzzleMix	76.6	94.3	95.5	72.3	49.5	68.4	55.0	59.5	63.5	0.467	0.627
SlideMix	**77.2**	**94.9**	**95.8**	**73.8**	**51.6**	**69.8**	**55.3**	**66.7**	**68.8**	**0.521**	**0.637**

Table 2. Impact of SlideMix on SOTA MIL Backbones Performance (Larger values in bold).

Method	PANDA		CAMELYON16		IMP-CRS-2024		TCGA-Lung		TCGA-BLCA		TCGA-UCEC		TCGA-CESC		TCGA-UCS		TCGA-UVM		TCGA-BRCA		TCGA-GBM	
	Acc. [%]		Acc. [%]		Acc. [%]		Acc. [%]		Acc. [%]		Acc. [%]		Acc. [%]		Acc. [%]		Acc. [%]		Corr.		Corr.	
	Base	Ours	Base	Ours	Base	Ours	Base	Ours	Base	Ours	Base	Ours	Base	Ours	Base	Ours	Base	Ours	Base	Ours	Base	Ours
SlideAve	66.7	**68.1**	69.6	**71.2**	92.3	**93.2**	**75.1**	74.6	51.1	**52.2**	68.3	**69.6**	52.6	52.6	58.3	**62.5**	**75.0**	71.9	0.529	**0.541**	0.539	**0.556**
SlideMax	62.5	**64.2**	**95.3**	94.6	93.1	**94.1**	**75.7**	75.1	53.8	**54.9**	69.2	**70.4**	47.4	**49.1**	41.7	**45.8**	31.2	**37.5**	**0.520**	0.514	0.463	**0.485**
ABMIL	75.9	**77.2**	93.7	**94.9**	95.2	**95.8**	71.6	**73.8**	48.4	**51.6**	67.5	**69.8**	54.4	**55.3**	58.3	**66.7**	62.5	**68.8**	0.449	**0.521**	0.614	**0.637**
CLAM	76.0	**76.8**	73.4	**74.7**	92.9	**93.7**	72.2	**73.2**	**50.0**	49.5	64.2	**65.8**	**54.4**	52.6	58.3	58.3	68.8	**71.9**	0.537	**0.551**	0.288	**0.308**
DSMIL	**78.5**	78.1	86.1	**87.3**	94.6	**95.4**	73.4	**74.3**	**56.8**	55.9	**71.7**	70.8	50.9	**52.6**	58.3	**63.3**	62.5	**65.6**	0.532	**0.548**	0.509	**0.528**
TransMIL	**76.2**	75.8	93.7	**94.3**	95.2	**95.9**	73.4	**74.1**	46.6	**48.4**	**69.2**	68.3	50.9	**52.6**	66.7	**70.8**	56.2	**59.4**	0.504	**0.519**	**0.648**	0.641
SETMIL	**78.2**	78.0	**94.1**	93.8	**95.6**	95.5	75.8	**76.0**	**57.3**	57.1	70.5	**70.6**	55.1	**55.2**	68.9	**69.1**	71.8	**72.0**	**0.562**	0.560	0.625	**0.627**
DTFD-MIL	**78.9**	78.7	**94.8**	94.5	95.9	**96.0**	**85.2**	85.0	58.1	**58.3**	**71.2**	71.0	**55.8**	55.5	**70.3**	70.1	72.5	**72.6**	**0.574**	0.572	**0.638**	0.635
GigaPath	77.9	**78.5**	84.5	**85.8**	94.9	**95.1**	75.1	**75.7**	52.2	**53.8**	68.3	**69.2**	50.9	**52.6**	50.0	**54.2**	56.2	**59.4**	**0.636**	0.629	0.618	**0.632**
MambaMIL	**78.6**	78.4	94.5	**94.7**	**95.7**	95.6	76.3	**76.5**	**58.9**	58.7	70.8	**71.0**	**54.7**	54.5	69.5	**69.8**	73.1	**73.3**	**0.568**	0.566	0.631	**0.633**

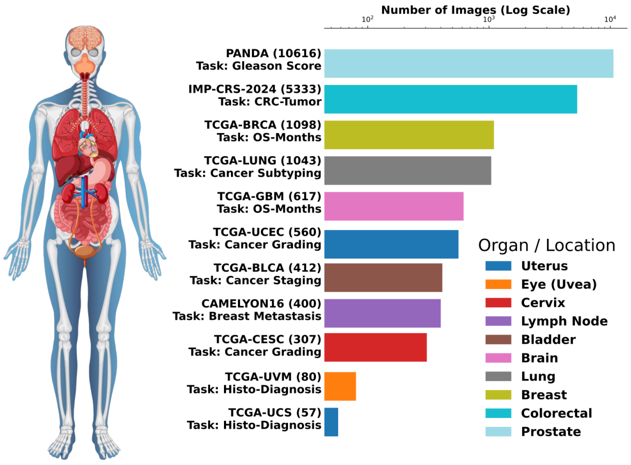

Figure 6. Summary of implemented datasets and tasks.

4.2. Implementation Details

The comparison in Tab. 1 uses ABMIL [15] as the slide-level backbone, while all backbones in Tab. 2 are initialized from scratch without pretraining. Models were trained for 100 epochs—20 for warmup and 80 for main training—using the Adam optimizer with an initial learning rate of 1×10^{-4} and a cosine decay schedule annealing to 1×10^{-6} (1% of the initial value). The UnPuzzle [18] framework was used for pre-processing, with each WSI divided into 224×224 tiles represented by Gigapath [33] feature embeddings. To minimize sampling variance, particularly for small datasets, 15 independent test inferences with different random tile samplings were performed per scenario, and predictions were aggregated. A batch size of 4 was maintained for all experiments, conducted on an Nvidia H100 GPU using Python 3.10.16, PyTorch 2.4.0, and CUDA 12.4.

4.3. Comparison with SOTA Methods

We benchmarked SlideMix against five augmentation strategies, including **CutMix** [36], **CutOut** [12], **MixUp** [37], **ResizeMix** [26], and **PuzzleMix** [16]. However, these methods operate on raw tiles or leverage saliency maps, making them directly incompatible with the MIL framework. To enable a fair comparison, we applied each augmentation at the raw image level and then compute tile embeddings from the resulting augmented images. For faithful implementation, we utilize the original code repositories for each baseline method. The baseline model was trained without any data augmentation.

Tab. 1 shows that SlideMix consistently achieves superior performance across all evaluated datasets. On CAMELYON16, while most augmentation methods demonstrate gains over the baseline (93.7%), SlideMix reaches the highest accuracy at 94.9% (+1.2%). The advantages of SlideMix are more pronounced on the more challenging datasets. On TCGA-Lung, SlideMix achieves 73.8% accuracy, representing a substantial improvement of 2.2% over the baseline (71.6%) and 0.9% over the next-best method, ResizeMix (72.9%). Similarly, on the PANDA dataset, SlideMix attains 77.2% accuracy, marking a 1.3% improvement over the baseline (75.9%). These results highlight two key findings: 1) mixing-based augmentation strategies generally outperform the baseline, confirming their effectiveness for WSI analysis; and 2) among these methods, SlideMix demonstrates the most consistent and substantial

Method	PANDA	CAMELYON16	IMP-CRS-2024	TCGA-Lung
	Acc. [%]	Acc. [%]	Acc. [%]	Acc. [%]
Baseline	75.9	93.7	95.2	71.6
Random	60.9	81.5	79.2	57.0
Linear	72.7	89.0	92.0	70.4
VLM/u	75.3	92.1	94.5	73.3
VLM	**77.2**	**94.9**	**95.8**	**73.8**

Table 3. Comparison between different soft labeling methods. VLM/u represents untuned VLM.

gains across diverse pathology tasks, including classification, staging, grading, and survival prediction.

To further verify the generalizability of SlideMix, we evaluate its performance against two baselines, **SlideAve** and **SlideMax** (SlideAve generates slide-level features via global average pooling across all tiles, whereas SlideMax employs global max pooling), and eight state-of-the-art MIL backbones: **ABMIL** [15], **DSMIL** [17], **CLAM** [20], **TransMIL** [28], **SETMIL** [43], **DTFD-MIL** [38], **GigaPath** [33], and **MambaMIL** [34]. As shown in Tab. 2, SlideMix generally improves performance across most backbones on the benchmark datasets. On the PANDA dataset, gains range from 0.6% to 1.7%, while CAMELYON16 shows improvements of 1.2% to 2.5%. Notably, substantial gains are observed for ABMIL on TCGA-Lung (from 71.6% to 73.8%) and IMP-CRC-2024 (from 95.2% to 95.8%). In a small number of settings, SlideMix's performance is marginally lower, likely due to dataset or model architecture differences, but overall it delivers clear and consistent gains across diverse models and datasets. The observed overall improvements across diverse backbones, from simple aggregation methods such as SlideMax to Transformer-based models such as TransMIL, suggest that SlideMix offers a generally model-agnostic enhancement strategy for computational pathology.

5. Discussion

5.1. Visualization Analysis

To examine the impact of different augmentation methods on model attention, we visualize Class Activation Maps (CAMs) using Grad-CAM [9]. We compared CAMs from the baseline model (no augmentation) with those from various augmentation methods. As shown in Fig. 7, SlideMix effectively guides the model to focus on task-relevant regions and delineate feature boundaries more clearly. Compared to the baseline and other methods, SlideMix enables richer WSI representations, enhancing cross-feature attention at both local (specific pathological regions) and global scales (contextualizing regions within the whole slide).

5.2. Soft Labeling Approach Analysis

We compared the performance of the baseline with four different labeling strategies on SlideMix: The **Random** randomly chooses a label from one of the two source WSIs. **Linear** calculates the label l_s as a weighted average of the original labels l_r, l_m, and the weight f corresponds to the content ratio of two WSIs S_r, S_m in the shuffled WSI S_s.

$$l_s = f l_r + (1-f) l_m \quad l_r, l_m, l_s \in \mathbb{R}^C \tag{9}$$

where C is the number of classes. **VLM/u** directly applies the untuned CONCH as the soft labeler, and **VLM**, which we adopt, utilizes CONCH with RAG support. Tab. 3 shows that: 1) Random fails to provide accurate labels, substantially limiting model performance compared with the baseline (-15.0% Avg. Acc.). 2) Linear provides relatively more accurate labels but still struggles under weak supervision, performing below the baseline (-2.5% Avg. Acc.). 3) VLM/u significantly improves overall performance and surpasses the baseline ($+0.4\%$ Avg. Acc.), though it slightly underperforms on IMP-CRS-2024 (-0.7% Acc.). 4) VLM achieves improvements on all tested datasets ($+2.4\%$ Avg. Acc.), demonstrating the effectiveness of our approach.

Method	PANDA	CAMELYON16	IMP-CRS-2024	TCGA-Lung
	Eff. [s]	Eff. [s]	Eff. [s]	Eff. [s]
Raw Image	23.1	512.4	357.7	411.4
Embedding	**0.31**	**8.2**	**5.3**	**7.2**
	Acc. [%]	Acc. [%]	Acc. [%]	Acc. [%]
Sequential	75.4	93.1	93.9	72.8
Local-box	74.9	92.5	94.1	72.3
Z-order	76.4	94.1	94.7	73.3
Random	**77.2**	**94.9**	**95.8**	**73.8**

Table 4. Performance comparison between different shuffling methods. Eff. (Efficiency) represents processing time per WSI.

5.3. Shuffling & Sampling Approach Analysis

We first compared the stage at which shuffling is applied. **Raw image** takes the raw WSIs, pre-processes them, shuffles the WSIs, then uses GigaPath to generate the shuffled tile embeddings on the shuffled WSI. **Embedding** takes both raw WSIs and embedded tiles, generates the shuffled coordinates based on raw WSIs, then organizes the embedded tiles based on the coordinates. Tab. 4 shows that operating directly on tile embeddings skips the redundant tile-embedding process, significantly reducing the processing time per WSI during augmentation (65.4× Avg. Eff.).

We then compared four sampling strategies. In MIL, only a fixed-size subset of tiles in a WSI is sent into the slide-level backbone for training efficiency, and the sampling strategy controls which tiles are in the subset. The baseline, **Sequential**, directly samples the tiles row by row; **Z-order** samples tiles in Z-order to preserve the inter-tile spatial information [25]; **Local-box** selects several central points, then samples tiles in a given radius around these points; **Random** randomly samples tiles over the whole WSI. Tab. 4 shows that: 1) Local-box fails to sample more

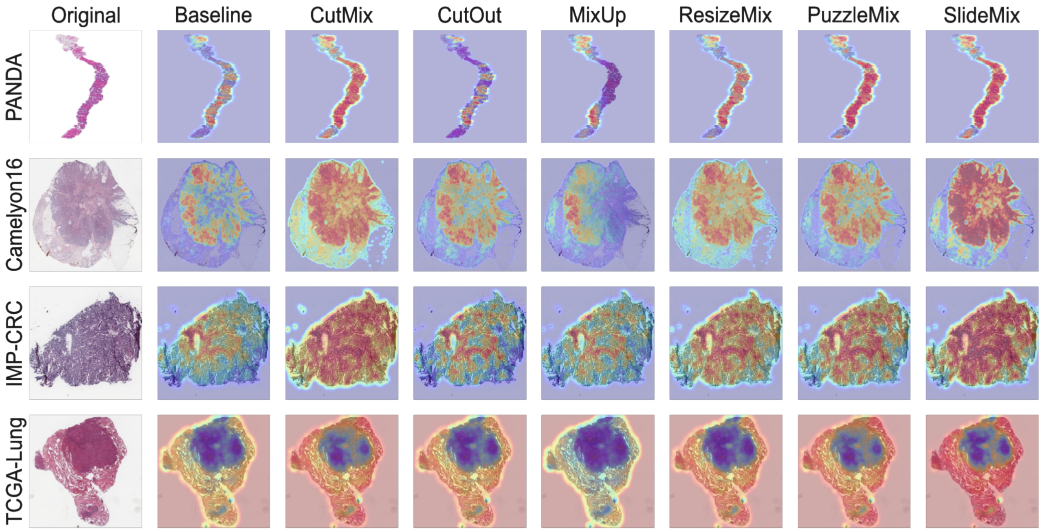

Figure 7. Grad-CAM visualization for ABMIL trained with different augmentation methods. Baseline is trained without any augmentation.

useful tiles compared with Sequential (-0.5% Avg. Acc.), as the feature integrity in shuffled WSIs are already disrupted, and local sampling doesn't help provide the useful spatial information. 2) Z-order performs slightly better than Sequential (+0.6% Avg. Acc.), as it is not severely affected by the shuffling process. 3) Random significantly surpasses Sequential (+1.5% Avg. Acc.), as it is inherently compatible with the shuffling process, and facilitates the model to learn cross-tile feature relationships.

5.4. Curriculum Learning Analysis

We first compared the performance of three loss-driven strategies. **Fixed** maintains constant difficulty parameters throughout the entire training process. **Loss-back** reduces the difficulty parameters when the validation loss l is lower than the performance threshold T_{loss}. **Loss-hold** maintains the current difficulty parameters when $l < T_{loss}$. Tab. 5 shows that: 1) Loss-back guides the model to learn features step by step. This dynamic process significantly outperforms Fixed (+0.3% Avg. Acc.). 2) Loss-hold further maintains the difficulty as a stricter teacher, further surpassing Fixed (+1.1% Avg. Acc.).

We evaluated each scheduler's contribution, where **All** activates all schedulers and **Set 1, 2, 3** disable the shuffle ratio, PCA similarity, and shuffle granularity schedulers, respectively. Tab. 5 shows that Set 3 has the worst performance (-2.9% Avg. Acc.), demonstrating the critical importance of shuffle granularity scheduling. Set 1 drops -2.3% Avg. Acc., highlighting the shuffle ratio scheduler's importance, while Set 2's slight decline (-0.6% Avg. Acc.) suggests room for improvement in PCA similarity scheduling. All schedulers contribute meaningfully to overall performance.

Method	PANDA	CAMELYON16	IMP-CRS-2024	TCGA-Lung
	Acc. [%]	Acc. [%]	Acc. [%]	Acc. [%]
Fixed	75.9	93.7	94.6	73.1
Loss-back	76.4	93.7	95.2	73.6
Loss-hold	**77.2**	**94.9**	**95.8**	**73.8**
Set 1	75.0	92.7	94.1	72.3
Set 2	76.6	94.2	95.4	73.3
Set 3	74.3	91.9	93.6	71.8
All	**77.2**	**94.9**	**95.8**	**73.8**

Table 5. Comparison between different loss-driven strategies and different schedulers. Shuffle ratio scheduler, PCA similarity threshold scheduler, and shuffle granularity scheduler are disabled respectively in set 1, 2, and 3.

6. Conclusion

We introduce SlideMix, a multimodal dynamic data augmentation framework that enhances WSI analysis. SlideMix employs VARS to adaptively select label-relevant regions for the in-place tile shuffling in ITS. This process is guided by CLF, which promotes progressive, multi-scale feature learning. Tabs. 1-5 and Fig. 1 show that SlideMix effectively mitigates three key challenges in WSI analysis: weak supervision, spatial heterogeneity, and cross-scale feature fusion. Extensive experiments across 11 benchmark datasets confirmed that SlideMix generally improves accuracy and generalization on diverse pathological tasks, outperforming existing SOTA methods like CutMix, MixUp, and PuzzleMix. By coupling multimodal reasoning with adaptive feature-level mixing, SlideMix provides a scalable and architecture-agnostic augmentation paradigm for computational pathology. Future work will focus on integrating generative priors to enhance sample realism and extending the SlideMix framework to multimodal clinical datasets that include genomics and radiology.

References

[1] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017. 5

[2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009. 3

[3] Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, et al. An introduction to vision-language modeling. *arXiv preprint arXiv:2405.17247*, 2024. 3

[4] Wouter Bulten, Kimmo Kartasalo, Po-Hsuan Cameron Chen, Peter Ström, Hans Pinckaers, Kunal Nagpal, Yuannan Cai, David F Steiner, Hester Van Boven, Robert Vink, et al. Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. *Nature medicine*, 28(1):154–163, 2022. 5

[5] Kathi Canese and Sarah Weis. Pubmed: the bibliographic database. *The NCBI handbook*, 2(1):2013, 2013. 4

[6] Di Chen, Yiwei Bai, Wenting Zhao, Sebastian Ament, John Gregoire, and Carla Gomes. Deep reasoning networks for unsupervised pattern de-mixing with constraint reasoning. In *International Conference on Machine Learning*, pages 1500–1509. PMLR, 2020. 3

[7] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Bowen Chen, Andrew Zhang, Daniel Shao, Andrew H Song, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 2024. 1, 3

[8] Sicheng Chen, Tianyi Zhang, Dankai Liao, Dandan Li, Low Chang Han, Yanqin Jiang, Yueming Jin, and Shangqing Lyu. Pathrwkv: Enabling whole slide prediction with recurrent-transformer. *arXiv preprint arXiv:2503.03199*, 2025. 2

[9] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, pages 1251–1258, 2017. 7

[10] Antonio Colaprico, Tiago C Silva, Catharina Olsen, Luciano Garofano, Claudia Cava, Davide Garolini, Thais S Sabedot, Tathiane M Malta, Stefano M Pagnotta, Isabella Castiglioni, et al. Tcgabiolinks: an r/bioconductor package for integrative analysis of tcga data. *Nucleic acids research*, 44(8):e71–e71, 2016. 5

[11] Cong Cong, Sidong Liu, Antonio Di Ieva, Maurice Pagnucco, Shlomo Berkovsky, and Yang Song. Colour adaptive generative networks for stain normalisation of histopathology images. *Medical Image Analysis*, 82:102580, 2022. 2, 3

[12] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 6

[13] Junsong Fan, Zhaoxiang Zhang, and Tieniu Tan. Employing multi-estimations for weakly-supervised semantic segmentation. In *European conference on computer vision*, pages 332–348. Springer, 2020. 1

[14] Matthew G Hanna, Anil Parwani, and Sahussapont Joseph Sirintrapun. Whole slide imaging: technology and applications. *Advances in Anatomic Pathology*, 27(4):251–259, 2020. 1

[15] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018. 2, 3, 6, 7

[16] Jang-Hyun Kim, Wonho Choo, and Hyun Oh Song. Puzzle mix: Exploiting saliency and local statistics for optimal mixup. In *International Conference on Machine Learning*, pages 5275–5285, 2020. 6

[17] Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328, 2021. 7

[18] Dankai Liao, Sicheng Chen, Nuwa Xi, Qiaochu Xue, Jieyu Li, Lingxuan Hou, Zeyu Liu, Chang Han Low, Yufeng Wu, Yiling Liu, Yanqin Jiang, Dandan Li, and Shangqing Lyu. Unpuzzle: A unified framework for pathology image analysis, 2025. 1, 6

[19] Haoneng Lin, Cheng Xu, and Jing Qin. Taming vision-language models for medical image analysis: A comprehensive review. *arXiv preprint arXiv:2506.18378*, 2025. 3

[20] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570, 2021. 7

[21] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature medicine*, 30(3):863–874, 2024. 5

[22] Andrzej Maćkiewicz and Waldemar Ratajczak. Principal components analysis (pca). *Computers & Geosciences*, 19 (3):303–342, 1993. 4

[23] Sara P Oliveira, Pedro C Neto, João Fraga, Diana Montezuma, Ana Monteiro, João Monteiro, Liliana Ribeiro, Sofia Gonçalves, Isabel M Pinto, and Jaime S Cardoso. Cad systems for colorectal cancer from wsi are still not ready for clinical acceptance. *Scientific Reports*, 11(1):14358, 2021. 5

[24] Pedro Osorio, Guillermo Jimenez-Perez, Javier Montalt-Tordera, Jens Hooge, Guillem Duran-Ballester, Shivam Singh, Moritz Radbruch, Ute Bach, Sabrina Schroeder, Krystyna Siudak, et al. Latent diffusion models with image-derived annotations for enhanced ai-assisted cancer diagnosis in histopathology. diagnostics 14 (13)(2024). issn: 2075-4418, 2024. 3

[25] Jie Peng, Jingxia Jiang, Yueliang Ying, Sukwon Yun, Qi Long, Yanyong Zhang, and Tianlong Chen. One leaf knows autumn: A piece of data-model facilitates efficient cancer

prognosis with histological and genomic modalities. In *Workshop on Large Language Models and Generative AI for Health at AAAI 2025*, 2025. 2, 3, 7

[26] Jie Qin, Jiemin Fang, Qian Zhang, Wenyu Liu, Xingang Wang, and Xinggang Wang. Resizemix: Mixing data with preserved object information and true labels. *arXiv preprint arXiv:2012.11101*, 2020. 6

[27] Adalberto Claudio Quiros, Roderick Murray-Smith, and Ke Yuan. Pathologygan: Learning deep representations of cancer tissue. arxiv 2019. *arXiv preprint arXiv:1907.02644*, 1907. 3

[28] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34:2136–2147, 2021. 1, 5, 7

[29] Chetan L Srinidhi, Ozan Ciga, and Anne L Martel. Deep neural network models for computational histopathology: A survey. *Medical image analysis*, 67:101813, 2021. 2

[30] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 4

[31] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Junzhou Huang, Wei Yang, and Xiao Han. Transpath: Transformer-based self-supervised learning for histopathological image classification. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 186–195, Cham, 2021. 1

[32] Yu Xin, Gorkem Can Ates, Kuang Gong, and Wei Shao. Med3dvlm: An efficient vision-language model for 3d medical image analysis. *arXiv preprint arXiv:2503.20047*, 2025. 3

[33] Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, pages 1–8, 2024. 6, 7

[34] Shu Yang, Yihui Wang, and Hao Chen. Mambamil: Enhancing long sequence modeling with sequence reordering in computational pathology. In *International conference on medical image computing and computer-assisted intervention*, pages 296–306. Springer, 2024. 7

[35] Nan Ying, Yanli Lei, Tianyi Zhang, Shangqing Lyu, Chunhui Li, Sicheng Chen, Zeyu Liu, Yu Zhao, and Guanglei Zhang. Cpia dataset: A comprehensive pathological image analysis dataset for self-supervised learning pre-training. *arXiv preprint arXiv:2310.17902*, 2023. 1

[36] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, pages 6023–6032, 2019. 6

[37] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 6

[38] Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E Coupland, and Yalin Zheng. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18802–18812, 2022. 7

[39] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5625–5644, 2024. 3

[40] Tianyi Zhang, Youdan Feng, Yunlu Feng, Yu Zhao, Yanli Lei, Nan Ying, Zhiling Yan, Yufang He, and Guanglei Zhang. Shuffle instances-based vision transformer for pancreatic cancer rose image classification. *arXiv preprint arXiv:2208.06833*, 2022. 1

[41] Tianyi Zhang, Shangqing Lyu, Yanli Lei, Sicheng Chen, Nan Ying, Yufang He, Yu Zhao, Yunlu Feng, Hwee Kuan Lee, and Guanglei Zhang. Puzzletuning: Explicitly bridge pathological and natural image with puzzles. *arXiv preprint arXiv:2311.06712*, 2023. 1

[42] Tianyi Zhang, Zhiling Yan, Chunhui Li, Nan Ying, Yanli Lei, Yunlu Feng, Yu Zhao, and Guanglei Zhang. Cellmix: A general instance relationship based method for data augmentation towards pathology image classification. *arXiv preprint arXiv:2301.11513*, 2023. 1, 2

[43] Yu Zhao, Zhenyu Lin, Kai Sun, Yidan Zhang, Junzhou Huang, Liansheng Wang, and Jianhua Yao. Setmil: spatial encoding transformer-based multiple instance learning for pathological image analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 66–76. Springer, 2022. 7