This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

ModalTune: Fine-Tuning Slide-Level Foundation Models with Multi-Modal Information for Multi-task Learning in Digital Pathology

Vishwesh Ramanathan[1,2*] Tony Xu[1,2*] Pushpak Pati[4] Faruk Ahmed[3] Maged Goubran[1,2]
Anne L. Martel[1,2]
[1]Sunnybrook Research Institute, Canada
[2]University of Toronto, Canada
[3]Google Research, USA
[4]Independent Researcher

Abstract

Prediction tasks in digital pathology are challenging due to the massive size of whole-slide images (WSIs) and the weak nature of training signals. Advances in computing, data availability, and self-supervised learning (SSL) have paved the way for slide-level foundation models (SLFMs) that can improve prediction tasks in low-data regimes. However, current methods under-utilize shared information between tasks and modalities. To overcome this challenge, we propose ModalTune, a novel fine-tuning framework which introduces the Modal Adapter to integrate new modalities without modifying SLFM weights. Additionally, we use large-language models (LLMs) to encode labels as text, capturing semantic relationships across multiple tasks and cancer types in a single training recipe. ModalTune achieves state-of-the-art (SOTA) results against both unimodal and multi-modal models across four cancer types, jointly improving survival and cancer subtype prediction while remaining competitive in pan-cancer settings. Additionally, we show ModalTune is generalizable to two out-of-distribution (OOD) datasets. To our knowledge, this is the first unified fine-tuning framework for multi-modal, multi-task, and pan-cancer modeling in digital pathology. Our code is publicly available at https://github.com/martellab-sri/ModalTune.

1. Introduction

Prediction tasks in histopathology are challenging due to the massive size and complexity of WSIs, which can contain critical diagnostic information for tumor staging, subtyping, and prognosis assessment. These predictive signals are often at the slide-level, requiring weakly-supervised learning approaches. Multiple Instance Learning (MIL) has become a key method for WSI analysis, dividing WSIs into patches, extracting features using pre-trained models, and aggregating them into slide-level representations through various pooling methods [9, 28, 54, 59, 61, 79]. With the development of large patch-level foundation models (FMs) [11, 13, 43, 69], feature extraction has improved, thereby enhancing the performance of MIL in slide-level tasks [7].

Recent advances in computational resources, data availability, and algorithm efficiency have led to the development of SLFMs, which involve training MIL aggregators ('slide encoders') using SSL techniques. These SSL techniques include masked image modeling [25, 72], contrastive learning with transcriptomics [29, 64], pathology reports [1, 17], and multiple stains [32], marking a shift from patch-level FMs to robust SLFMs. Yet, many domain-specific variations may still not be captured during pre-training, such as differences in tissue preparation across clinical centers, OOD morphological characteristics, and the presence of complex or rare diseases. This makes fine-tuning of SLFMs a more effective strategy for deployment on new datasets.

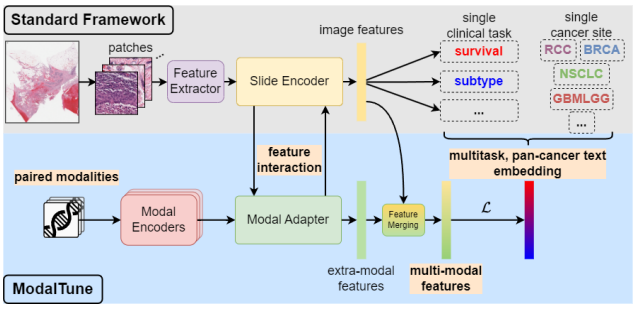

Figure 1. Unlike standard fine-tuning approaches that focus on a single task, modality, and cancer site, our novel fine-tuning framework, ModalTune, leverages shared information across multiple tasks, modalities, and cancer sites to interface with SLFMs.

*Equal contributions

Current frameworks for deploying SLFMs are limited to fine-tuning them on a single task and cancer site while taking only WSIs as input (Fig. 1). They fail to fully leverage rich datasets, which may include other modalities like clinical records, reports, and genomics, as well as extensive slide annotations like cancer subtypes, tumor staging, prognosis, and biomarker presence. Existing frameworks face two key limitations when applied in this setting: (1) Existing SLFMs use only WSIs as input during inference, overlooking other modalities. While these models may have incorporated other modalities for training, there are significant benefits to leveraging them during inference to capitalize on multi-modal insights [10, 31]. (2) Current frameworks often emphasize single-task learning, overlooking interconnected multi-task information that could enhance SLFMs' fine-tuning performance and generalization [14]. These challenges result in suboptimally fine-tuned SLFMs.

To address these limitations, we propose two enhancements that transform SLFMs into multi-modal and multi-task learners while preserving their foundational knowledge: (1) We introduce the *Modal Adapter*. Inspired by recent work in NLP and natural images [12, 51], this module interfaces with slide-encoders in histopathology to integrate information from additional modalities without altering the primary model's weights. (2) We introduce *text-based multi-task learning*. We convert downstream tasks (*e.g*. subtype classification, risk prediction) into text and use LLMs to represent them as text vectors in a unified embedding space. We tune our models to predict these text embeddings using a single loss function, regardless of the task type or number, enabling the model to leverage semantic relationships between entities within and across tasks.

Our overall method called *ModalTune* is a fine-tuning framework designed to fully leverage the potential of SLFMs and pathology datasets. Our main contributions are:

- **Modal Adapters**: We propose Modal Adapters to continually inject multi-modal information into slide encoders without altering their weights. It can interface with any transformer-based slide encoder and dynamically scale to new modalities.
- **Multi-Task Learning**: We simplify multi-task learning by mapping tasks into a shared embedding space using LLMs, leveraging complementary semantic and inter-task relationships. Our formulation optimizes a single loss function regardless of the number or type of tasks.
- **Pan-Cancer Fine-Tuning**: Our framework integrates multiple cancer sites, resulting in a single resource-efficient, pan-cancer model that generalizes across tasks and sites. We show this approach enables smaller datasets to leverage key insights from larger ones.

To the best of our knowledge, we propose the first fine-tuning framework for unified approach for multi-modal, multi-task, and pan-cancer modeling in digital pathology.

We validate our approach on four TCGA datasets [63]: Breast Invasive Carcinoma (BRCA), Non-small Cell Lung Carcinoma (NSCLC), Glioblastoma Multiforme and Low-Grade Glioma (GBMLGG), and Renal Cell Carcinoma (RCC), covering 4,099 slides and 3,418 patients. Using Prov-Gigapath [72] (referred to as Gigapath) as our SLFM, our model outperforms uni-modal and multi-modal baselines in cancer subtype and survival prediction, achieving higher balanced accuracy and C-indices than the fine-tuned uni-modal Gigapath (+4.6%, +9.6%), its multi-modal counterpart (+2.8%, +7.0%), and the best baseline (+1.4%, +1.9%). We further tested our model on two smaller OOD datasets: Colon and Rectum Adenocarcinoma (COAD-READ) and Bladder Urothelial Carcinoma (BLCA), where it retained generalizability, performed competitively with the supervised baseline and outperformed its fine-tuned counterpart by a large margin (+25%, +8.3%).

2. Related Work

This section reviews related works on WSI-text alignment, multi-modal and multi-task methods in digital pathology.

WSI-Text Alignment: Recent vision-language models (e.g. CLIP [52], CoCa [74], Llava [41], and BLIP-2 [38]) have advanced text alignment at the patch-level in digital pathology [43, 44, 80]. More recently, studies have aligned WSIs with slide-level diagnostic reports, with PRISM [57] using attention mechanisms and CoCa [74], and Gigapath [72] leveraging masked autoencoder pre-training [25] and contrastive learning on pathology reports. TITAN [17] enhances WSI-text alignment by first aligning slide embeddings with synthetic region-level captions, then refining them using slide-level pathology reports for improved representation. PathAlign [1] under the BLIP-2 [38] framework integrates slide encoder with frozen LLMs for report generation, WSI classification, and text retrieval.

While our work shares similarities with contrastive learning-based WSI-text alignment, it introduces key differences. Existing methods rely on large-scale datasets, pathology reports, and large batch sizes, limiting their adaptability. In contrast, our approach aligns more closely with typical MIL frameworks, enabling fine-tuning in a single-batch setting using widely available synoptic data, hence making it adaptable to any dataset.

Multi-Modal Models: Multiple studies in digital pathology show that integrating WSIs with clinical data—such as RNA transcriptomics [8, 10, 31, 33], copy number alterations (CNA) [10, 16], MRI scans [50], or clinical captions [80]—captures complementary biological and contextual insights, improving prediction over single-modality models. Multi-modal fusion in digital pathology employ various strategies to integrate heterogeneous data. Early fusion combines raw input features before processing, with initial methods using concatenation or pooling [27]. Inspired by

advances in Visual Question Answering field, recent approaches leverage cross-attention mechanisms [10, 31, 80]. Late fusion aggregates predictions from modality-specific models using methods like concatenation [18, 46], Kronecker products [8], or transformer-based fusion [73].

Our Modal Adapters, inspired by ViT Adapters [12], differ by performing cross-attention between modalities at multiple hidden layers, followed by summation-based fusion to generate the final feature representations.

Multi-Task Learning: Multi-task learning (MTL) optimizes related tasks from the same input by leveraging shared representations, enhancing data efficiency, generalizability, and reducing overfitting [14], making it especially useful in digital pathology which has co-related tasks like cancer subtyping, survival, and biomarker prediction. Early work at the patch level by Li et al. [39] used histopathology patches for breast cancer classification and grading, while Simon et al. [24] proposed a unified model for cell and tissue segmentation and classification. At the WSI-level, Weng et al. [71] introduced MTL for predicting multiple slide-level metadata, Gao et al. [20] developed a framework for cancer region detection and subtyping, and Zhao et al. [78] proposed a multi-task graph transformer for tumor typing and staging. More recently, Shao et al. [58] explored MTL for cancer subtpying, survival prediction, and transcriptomic profiling of gene TP53.

Unlike prior works that optimize multiple task-specific loss functions, our approach uses LLMs to unify tasks into a single representation, requiring only a single loss function for training. This offers greater simplicity and flexibility, making it easier to adapt to diverse or additional task sets.

3. Method

We introduce ModalTune, illustrated in Fig. 2, starting with conventional pretrained slide encoder (Sec. 3.1), and feature extraction from additional paired modalities (Sec. 3.2). We then present the Modal Adapter (Sec. 3.3) that injects additional modality features into the slide encoder. Finally, we describe the text embedding process (Sec. 3.4) for unifying tasks and cancer sites into a single embedding space and explain the overall training objective (Sec. 3.5).

3.1. Slide Encoder

We build on the SLFM Gigapath [72], chosen for its strong performance, computational efficiency, and publicly available pre-trained weights. Gigapath effectively models slide embeddings by leveraging all patches and scales well in scenarios involving multiple slides with a large number of patches, enabling analysis without excessive overhead. This balance of efficiency and coverage makes it a practical choice for our study, even though more recent SLFMs may achieve stronger benchmark performance. Importantly, our framework remains fully adaptable to any transformer-based SLFM. For experiments on ModalTune applied to recent SLFM like TITAN [17], please refer to Supp. Sec. 12.

Slide encoders extract WSI features using a two-step process. First, a WSI $\mathbf{X}_{img}$ is divided into N_{img} patches, and Gigapath's pre-trained patch encoder extracts D_{img} dimensional features, resulting in $\hat{\mathbf{X}}_{img} \in \mathbb{R}^{N_{img} \times D_{img}}$. The pre-trained Gigapath model, an L-layer LongNet [68], facilitates interactions among patch embeddings. Before further processing, these patch features pass through a linear layer to produce $\mathcal{F}^0_{img} \in \mathbb{R}^{N_{img} \times D}$, where D is the LongNet embedding size. For ease of notation in Sec. 3.3, the L LongNet layers are divided into B blocks, each with L/B layers. Following the Gigapath model, we use the frozen pre-trained `[CLS]` token from the slide encoder to aggregate patch embeddings into a single representation. Thus, after passing through the i-th LongNet block, the patch embeddings are $\mathcal{F}^i_{img} \in \mathbb{R}^{(N_{img}+1) \times D}$, for $i \in \{1, 2, ..., B\}$.

3.2. Modal Encoders

Modal encoders are employed to extract features from each additional modality, which will be used to inform the pre-trained slide encoder. We define M generic new modalities, denoted as m_h for $h \in \{1, 2, \ldots, M\}$, with input data $\mathbf{X}_{m_h}$ paired with the WSI $\mathbf{X}_{img}$. Each modality's input is independently encoded by the modal encoders, producing a feature set $\mathcal{F}^0_{m_h} \in \mathbb{R}^{N_{m_h} \times D}$, where N_{m_h} is the token length for each modality. The final multi-modal feature set is formed by concatenating features from all modalities, resulting in $\mathcal{F}^0_{mm} \in \mathbb{R}^{(N_{m_1}+N_{m_2}+...+N_{m_M}) \times D}$.

3.2.1. Transcriptomics Encoder

Bulk transcriptomics data is increasingly available and provides complementary insights on cancer biology to WSIs [8, 10, 31], allowing it to effectively predict patient survival [5, 19]. Thus, we consider it our primary additional modality in this work. Other modalities, such as clinical data (see Supp. Sec. 13), additional "omics," imaging techniques, or domain-specific priors, can in principle be integrated by concatenating their extracted feature tokens. However, further exploration of these extensions is left for future work.

The transcriptomics encoder compresses bulk transcriptomics data, $\mathbf{X}_t \in \mathbb{R}^{N_g}$, containing N_g genes into a set of tokens. We begin by dividing the genes into N_{gp} gene pathways, which are encoded using a sparse multi-layer perceptron (S-MLP) [19] into pathway features, $\tilde{\mathcal{F}}_{gp} \in \mathbb{R}^{N_{gp} \times D_{gp}}$, as described in [31]. To merge intra- and inter-pathway features, we use the MLP-mixer [62], encoding pathway features to produce $\mathcal{F}_{gp} \in \mathbb{R}^{N_{gp} \times D}$. Finally, a linear projection reduces the number of pathways to N_t, enhancing downstream tractability. The encoder finally outputs a compressed set of pathway tokens, $\mathcal{F}^0_t \in \mathbb{R}^{N_t \times D}$.

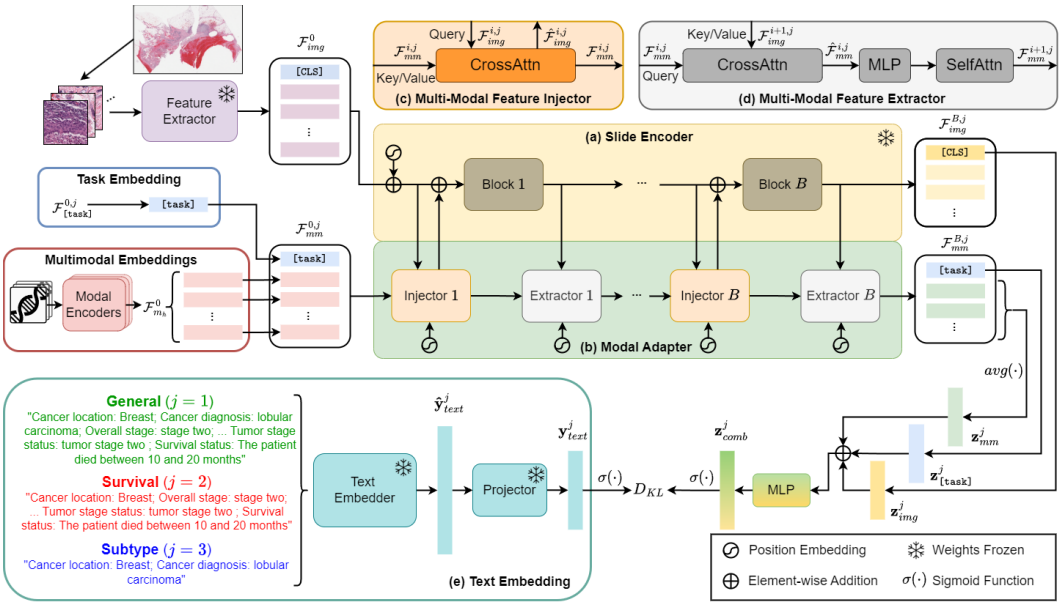

Figure 2. Overview of ModalTune. Built on a frozen pre-trained slide encoder (**a**), our Modal Adapter (**b**) integrates extra-modal features via the Multi-Modal Feature Injector (**c**) and Extractor (**d**). A text embedding module (**e**) unifies multiple tasks and cancer sites, while the enriched image, task, and modal embeddings are fused for training, ensuring a robust multi-modal, multi-task representation.

3.3. Modal Adapter

To integrate multi-modal features while taking advantage of pre-trained slide encoders, we introduce the Modal Adapter. Our adapter introduces informative cross-modality features into slide encoders, drawing inspiration from the ViT-Adapter [12]. Specifically, it merges compressed pathway features from the transcriptomics encoder with image features from the slide encoder, enriching the network with meaningful multi-modal representations.

3.3.1. Task Prompt

The Modal Adapter requires distinct feature sets tailored to each downstream task. To facilitate feature interactions and attention operations, we introduce a learnable `[task]` prompt. This consists of a set of learnable vectors, $\mathcal{F}^{0,j}_{\texttt{[task]}} \in \mathbb{R}^{1\times D}$, for each downstream task $j \in \{1, 2, \ldots, T\}$, capturing task-specific biases across T tasks. This prompt is concatenated with N_t modal tokens derived from compressed pathways in our approach, to create the complete set of multi-modal features: $\mathcal{F}^{0,j}_{mm} \in \mathbb{R}^{(N_t+1)\times D}$.

3.3.2. Multi-modal Feature Interaction

Feature interaction in the Modal Adapter is governed by *Multi-modal Feature Injector* modules, which inject multi-modal features into image features, and *Multi-modal Feature Extractor* modules, which extract image features to inform the features from extra modalities.

Multi-Modal Feature Injectors are formulated as standard cross-attention, CrossAttn(·) [65], with query being the WSI features, $\mathcal{F}^{i,j}_{img} \in \mathbb{R}^{(N_{img}+1)\times D}$, and key and value being the multi-modal features, $\mathcal{F}^{i,j}_{mm} \in \mathbb{R}^{(N_t+1)\times D}$. The Injector for the i-th LongNet block and j-th task is given as,

$$\hat{\mathcal{F}}^{i,j}_{img} = \mathcal{F}^{i,j}_{img} + \gamma^i \text{CrossAttn}(\text{LN}(\mathcal{F}^{i,j}_{img}), \text{LN}(\mathcal{F}^{i,j}_{mm})) \quad (1)$$

where, LN(·) is LayerNorm [4], and $\gamma^i \in \mathbb{R}^D$ is a learnable vector controlling how the original image features are adjusted by the cross-attention layer. Based on [12], we initialize γ^i with **0** to allow multi-modal features to be *gradually* introduced into the pre-trained slide encoder during training. The image features from the next LongNet block, $\mathcal{F}^{i+1,j}_{img}$, are obtained by passing the post-injection image features, $\hat{\mathcal{F}}^{i,j}_{img}$ through the i-th LongNet block.

Multi-modal Feature Extractors introduce image features *back into* multi-modal features. It is implemented using cross-attention, followed by an MLP and a self-attention operation SelfAttn(·) to model interactions between different modalities and task tokens. Here, the query is $\mathcal{F}^{i,j}_{mm}$, and the key and value are the image features after passing through a LongNet block, $\mathcal{F}^{i+1,j}_{img}$, as shown in Eq. (2) and Eq. (3).

$$\hat{\mathcal{F}}^{i,j}_{mm} = \mathcal{F}^{i,j}_{mm} + \text{CrossAttn}(\text{LN}(\mathcal{F}^{i,j}_{mm}), \text{LN}(\mathcal{F}^{i+1,j}_{img})) \quad (2)$$

$$\mathcal{F}^{i+1,j}_{mm} = \text{SelfAttn}\left(\hat{\mathcal{F}}^{i,j}_{mm} + \text{MLP}(\text{LN}(\hat{\mathcal{F}}^{i,j}_{mm}))\right) \quad (3)$$

3.3.3. Output Feature Representation

We extract the representation from the final LongNet block, $\mathcal{F}^{B,j}_{img} \in \mathbb{R}^{(N_{img}+1)\times D}$, using the `[CLS]` token to get the *image* feature, $\mathbf{z}^j_{img} \in \mathbb{R}^D$. We use the final multi-modal

features $\mathcal{F}_{mm}^{B,j} \in \mathbb{R}^{(N_t+1)\times D}$, indexed by the `[task]` token, to obtain the task-specific vector, $\mathbf{z}_{[\texttt{task}]}^{j} \in \mathbb{R}^{D}$. We also average the multi-modal features across the N_t compressed pathway tokens to get a modality-specific feature, $\mathbf{z}_{mm}^{j}$. Here, $\mathbf{z}_{[\texttt{task}]}^{j}$ encapsulates task-specific information enriched with image and extra-modality features, while $\mathbf{z}_{mm}^{j}$ captures modality-specific information contextualized by image and task features. We average modality features to form $\mathbf{z}_{mm}^{j}$ because the $\mathbf{z}_{[\texttt{task}]}^{j}$ primarily captures enriched task information, making it inadequate to represent modality features independently. Therefore, excluding the modality representation $\mathbf{z}_{mm}^{j}$ would lead to an incomplete feature set. The final output representation, $\mathbf{z}_{comb}^{j} \in \mathbb{R}^{D_{final}}$ combining all features tailored for the j-th task is produced as,

$$\mathbf{z}_{comb}^{j} = \text{MLP}(\text{LN}(\mathbf{z}_{img}^{j} + \mathbf{z}_{mm}^{j} + \mathbf{z}_{[\texttt{task}]}^{j})) \quad (4)$$

This vector can be used alongside classical machine learning methods on downstream tasks. Importantly, the Modal Adapter does not update the slide encoder weights, taking better advantage of pre-trained SLFMs and reducing the computational cost of our methodology.

3.4. Text Embedding

We encode tasks using text to unify tasks, datasets, and cancer sites into a single embedding space, eliminating the need for complex loss functions. We generate text embeddings by converting clinical tabular data into text (Fig. 2) and processing it with a pre-trained text encoder. Here, we use the text encoder from CONCH [43], pre-trained on histology captions. Since different tasks require unique feature interactions (Sec. 3.3.1), we create a 'general' task ($j = 1$) text embedding with all clinically relevant information and 'task-specific' ($j > 1$) embeddings tailored to each downstream task. The text for the j-th task is encoded into a vector, $\hat{\mathbf{y}}_{text}^{j} \in \mathbb{R}^{D_{text}}$, and passed through a fixed, randomly-initialized *Projector* to produce the final text embedding, $\mathbf{y}_{text}^{j} \in \mathbb{R}^{D_{final}}$. The Supplementary provides further details on text construction (Sec. 9) and analysis (Sec. 10).

3.5. Training Objective

We aim to replicate the text embedding distribution to capture both task-specific information and semantic relationships. To achieve this, we use Kullback–Leibler (KL) divergence, following [40], who demonstrated its advantages over CLIP [53] and other distillation losses. ModalTune is optimized to align the multi-modal embedding $\mathbf{z}_{comb}^{j}$ for task j with its corresponding text embedding $\mathbf{y}_{text}^{j}$. The alignment is achieved by minimizing the KL divergence loss, averaged across all tasks, as shown in Eq. (5), where $\sigma(\cdot)$ denotes softmax function.

$$\mathcal{L} = \frac{1}{T}\sum_{j=1}^{T} D_{KL}\left[\sigma\left(\frac{\mathbf{z}_{comb}^{j}}{||\mathbf{z}_{comb}^{j}||_2}\right) \Big|\Big| \sigma\left(\frac{\mathbf{y}_{text}^{j}}{||\mathbf{y}_{text}^{j}||_2}\right)\right] \quad (5)$$

	BRCA	GBMLGG	NSCLC	RCC	COADREAD	BLCA
Train cases	747	401	667	574	273	256
Val cases	113	71	100	102	38	46
Test cases	188	118	167	170	64	76
# events	146	200	362	219	85	171
# censorships	902	390	572	627	290	207
Rare-set	111	0	103	0	58	0
Class 0	746	102	378	271	248	312
Class 1	191	488	453	510	69	66
Class 2	-	-	-	65	-	-

Table 1. Dataset statistics. Note: 'Rare-set' denotes cancer subtypes not considered for training due to their relative rarity.

4. Experiments and Results

4.1. Datasets

We utilized four TCGA datasets [63]—BRCA, NSCLC, GBMLGG, and RCC—to train and evaluate ModalTune for cancer subtype and survival prediction, covering 4,099 slides and 3,418 patients. COADREAD and BLCA datasets were used as OOD data to test model generalizability. Data was split at patient-level into [68%, 12%, 20%] for training, validation, and testing, stratified by cancer subtypes (Tab. 1). Transcriptomics data were sourced from the Xena database [23] and preprocessed similar to SurvPath [31], resulting in 331 pathways with 4,987 genes. For subtyping, subtypes were grouped into broader categories based on OncoTree code definitions [37], with rare subtypes (< 25 cases) assigned to a 'Rare-set' used only for training with text embeddings. Baseline models cannot use this class due to insufficient occurrences for effective label-based training. Exact groupings are detailed in the Supp. Sec. 8. For each patient, we extracted 256×256 patches from WSI foreground regions (via thresholding in HSV colorspace) at $0.5\mu m/pixel$. Since outputs are generated on a patient-wise basis, patches from multiple WSIs were concatenated into a single bag. Features of size $D_{img} = 1536$ were extracted from each patch using Gigapath's DINOv2 [47] pre-trained feature extractor.

4.2. Baseline Setup

Baselines are trained in a single-task, single-cancer manner for classification and survival tasks. 'WSI' baselines use only WSIs, including popular models like ABMIL [28] and TransMIL [59]. We also assess the standalone SLFM, Gigapath [72], with full end-to-end tuning. 'Genomics' baselines rely on bulk transcriptomics, including MLP, SNN [35], and S-MLP [19]. We also test ModalTune's transcriptomics encoder, integrating an MLP-mixer [62] with S-MLP. 'Multi-modal' baselines combine image and genomic data, including cross-attention-based early fusion methods like MCAT [10] and SurvPath [31]. We also evaluate late fusion methods, merging image-only baselines with MLP-mixers via concatenation [46] and Kronecker product [8].

4.3. ModalTune Setup

We configure ModalTune for $T = 3$ tasks: general, survival, and subtype prediction. Gene pathways are linearly projected to $N_t = 64$ dimensions for tractability, and the final output feature dimension is $D_{final} = 256$. Training runs for 30 epochs with a 0.0001 learning rate and batch size of 1. Additional details are in Supp. Sec. 7.

After training, we evaluate its feature quality for classification and survival prediction. Using the General ($j = 1$) task prompt, we extract features from both training and test datasets, selecting the best training epoch based on validation classification performance. For final alignment, we use linear probing (LP), fitting a linear classifier [49] for cancer subtypes and a Cox Proportional Hazards (CPH) model [15] for survival prediction. All ModalTune results are *multitask*, utilizing a single feature set for both tasks. 'ModalTune' experiments are trained per cancer site and applied to downstream tasks within the same site, while 'ModalTune Pan-Cancer' pools all cancer sites during training and fits separate classifiers or CPH models per site for prediction.

4.4. Generalization Study Setup

To assess the generalizability of our pipeline, we apply fine-tuned models to unseen TCGA datasets: COADREAD and BLCA. 'ModalTune Pan-Cancer' applies the trained in-domain model to extract features on OOD datasets. 'ModalTune' extracts features using the model trained on BRCA. 'Gigapath Surv./Cls. (cat)' applies the respective single-task models directly tuned on BRCA and removes the final prediction layer to extract WSI features. We perform linear probing after feature extraction for all these experiments. The final 'Gigapath Sup. (cat)' experiment uses *supervised learning* to directly tune the multi-modal Gigapath baseline on both tasks on the train sets of the OOD datasets.

4.5. Quantitative Analysis

To address class imbalances in subtype classification, we report balanced accuracy, while survival prediction is evaluated using C-indices. Tab. 2 presents subtype classification and survival prediction results, showing the mean and standard deviation across three random seed runs. We report ablation studies on key design choices, including Modal Adapters, text embeddings, text encoders, task prompts, and Projectors in the Supp. Sec. 11.

Overall Results: On average, we found ModalTune surpassed all other baselines on subtype classification and survival prediction tasks on multiple cancer sites, with $+1.4\%$ and $+1.9\%$ overall improvement, respectively, over the best baseline. ModalTune also greatly improved over the standard image-only SLFM framework, Gigapath (Tuned) ($+4.6\%$ in subtype, $+9.6\%$ in survival prediction). These large improvements can be partially attributed to the introduction of the additional bulk transcriptomics modality. However, a fully-tuned Gigapath model with a late fusion of transcriptomics features (Gigapath (cat) and (KP)) still performs 2.8% worse on classification and 7.0% worse on survival prediction compared to ModalTune, indicating the benefits of our overall framework. Though TITAN is a stronger standalone model, applying ModalTune with it yielded similar trends (supp. Sec. 12).

Importance of Robust Multi-modal Interactions: The best modality for prediction varied by task and cancer site. For instance, WSI-baselines outperformed Genomics-baselines in BRCA subtyping, while the reverse was observed for RCC survival prediction. Combining modalities generally enhanced performance, as seen in GBMLGG, where adding bulk transcriptomics led to near-perfect subtyping in multi-modal models, with a +7.4% improvement from single to multi-modal Gigapath. Further investigation using a linear classifier on transcriptomics (test balanced accuracy: 0.99) identified four genes—CDC26, IDH1, IL13RA2, and IL8—explaining 97.5% of accuracy, with IDH1 being a well-documented biomarker [66], potentially trivializing classification. Thus, multi-modal results on GBMLGG serve as a "sanity check" for the transcriptomics signal, which fusion models—except MCAT—effectively retained. However, naïve fusion of multi-modal features was not always effective, especially for survival prediction, where the best multi-modal TransMIL model (KP) dropped 1.5% compared to Gene Mixer, and Gigapath (cat) dropped 2.5%. These findings emphasize the need for a *robust mixing strategy* when integrating multi-modal features.

Generalization Study: As shown in Tab. 3, the fully tuned Gigapath (cat) models severely overfit on BRCA, resulting in poor performance on both subtype and survival prediction tasks. These models fail to extract meaningful features for subtyping, defaulting to predicting the most common class (balanced accuracy of $\sim$0.5). We also observe that ModalTune consistently outperforms Gigapath LP across both tasks and cancer sites. Furthermore, ModalTune Pan-Cancer demonstrated superior generalizability, performing only 2.4% worse than a *fully-supervised* Gigapath (cat) network in subtype prediction and actually outperforming it by 1.2% in survival prediction. ModalTune's ability to extract highly relevant features without exposure to OOD datasets during fine-tuning underscores its capacity to maintain the generalizability of SLFMs. This result is particularly promising for the downstream application of ModalTune to smaller unseen datasets.

Pan-Cancer Results: As shown in Tab. 2, training a single ModalTune Pan-Cancer model by merging cancer sites did not outperform training models independently for each site on in-domain datasets, with a decrease of 2.8% in classification and 3.0% in survival prediction. This aligns with findings in [30], where a specialized model outperformed the pan-cancer model for molecular status prediction. How-

		BRCA	GBMLGG	NSCLC	RCC	Overall
		Cancer Subtype Prediction				
WSI	**Gigapath LP** [72]	0.612	0.900	0.821	0.851	0.796
	ABMIL [28]	0.853 ± 0.015	0.931 ± 0.040	0.920 ± 0.012	0.921 ± 0.017	0.906
	TransMIL [59]	0.828 ± 0.011	0.978 ± 0.012	0.934 ± 0.007	0.918 ± 0.016	0.915
	Gigapath (Tuned) [72]	0.860 ± 0.013	0.931 ± 0.029	0.916 ± 0.012	0.939 ± 0.016	0.912
Genomics	**MLP**	0.752 ± 0.032	$\underline{0.998} \pm 0.002$	0.926 ± 0.007	0.883 ± 0.016	0.890
	SNN [35]	0.753 ± 0.019	0.991 ± 0.002	0.926 ± 0.001	0.902 ± 0.008	0.893
	S-MLP [19]	0.839 ± 0.015	$\mathbf{1.000} \pm 0.000$	0.941 ± 0.005	0.890 ± 0.003	0.917
	Gene Mixer	0.840 ± 0.014	$\mathbf{1.000} \pm 0.000$	0.932 ± 0.005	0.898 ± 0.036	0.917
Multi-modal	**MCAT** [10]	0.875 ± 0.014	0.973 ± 0.005	0.921 ± 0.006	0.933 ± 0.028	0.925
	SurvPath [31]	0.858 ± 0.028	0.995 ± 0.007	0.932 ± 0.007	0.936 ± 0.030	0.930
	ABMIL (cat) [28]	0.861 ± 0.010	$\mathbf{1.000} \pm 0.000$	0.951 ± 0.003	0.937 ± 0.028	0.937
	ABMIL (KP) [28]	$\underline{0.884} \pm 0.010$	$\underline{0.998} \pm 0.002$	0.939 ± 0.005	0.945 ± 0.031	$\underline{0.941}$
	TransMIL (cat) [59]	0.847 ± 0.027	$\mathbf{1.000} \pm 0.000$	0.950 ± 0.004	$\underline{0.957} \pm 0.020$	0.939
	TransMIL (KP) [59]	0.868 ± 0.017	$\mathbf{1.000} \pm 0.000$	0.946 ± 0.005	0.930 ± 0.011	0.936
	Gigapath (cat) [72]	0.850 ± 0.017	$\underline{0.998} \pm 0.002$	0.924 ± 0.011	0.926 ± 0.034	0.925
	Gigapath (KP) [72]	0.821 ± 0.020	$\mathbf{1.000} \pm 0.000$	$\mathbf{0.963} \pm 0.003$	0.927 ± 0.018	0.928
	ModalTune (Ours)	$\mathbf{0.899} \pm 0.026$	$\mathbf{1.000} \pm 0.000$	0.956 ± 0.010	$\mathbf{0.959} \pm 0.003$	**0.954**
	ModalTune Pan-Cancer (Ours)	0.858 ± 0.001	0.990 ± 0.009	$\underline{0.958} \pm 0.004$	0.902 ± 0.033	0.927
		Survival Prediction				
WSI	**Gigapath LP** [72]	0.647	0.795	0.562	0.679	0.671
	ABMIL [28]	0.712 ± 0.004	0.854 ± 0.004	0.582 ± 0.002	0.670 ± 0.004	0.704
	TransMIL [59]	0.742 ± 0.015	0.868 ± 0.009	0.586 ± 0.011	0.676 ± 0.003	0.718
	Gigapath (Tuned) [72]	0.680 ± 0.024	0.824 ± 0.017	0.546 ± 0.005	0.685 ± 0.012	0.684
Genomics	**MLP**	0.629 ± 0.039	0.884 ± 0.004	0.542 ± 0.013	0.720 ± 0.002	0.694
	SNN [35]	0.519 ± 0.051	0.883 ± 0.010	0.508 ± 0.012	0.723 ± 0.008	0.659
	S-MLP [19]	0.749 ± 0.030	0.887 ± 0.002	0.571 ± 0.011	$\underline{0.735} \pm 0.003$	$\underline{0.736}$
	Gene Mixer	$\underline{0.762} \pm 0.049$	0.870 ± 0.007	0.556 ± 0.033	0.690 ± 0.009	0.719
Multi-modal	**MCAT** [10]	0.673 ± 0.044	0.880 ± 0.006	0.592 ± 0.002	0.697 ± 0.004	0.710
	SurvPath [31]	0.741 ± 0.031	$\underline{0.895} \pm 0.003$	$\mathbf{0.613} \pm 0.010$	0.677 ± 0.011	0.732
	ABMIL (cat) [28]	0.736 ± 0.007	$\mathbf{0.896} \pm 0.004$	0.605 ± 0.011	0.690 ± 0.004	0.732
	ABMIL (KP) [28]	0.685 ± 0.008	0.895 ± 0.002	0.594 ± 0.010	0.699 ± 0.003	0.718
	TransMIL (cat) [59]	0.666 ± 0.018	0.872 ± 0.030	0.595 ± 0.003	0.689 ± 0.004	0.705
	TransMIL (KP) [59]	0.689 ± 0.014	0.889 ± 0.003	0.541 ± 0.023	0.713 ± 0.003	0.708
	Gigapath (cat) [72]	0.678 ± 0.004	0.861 ± 0.023	0.573 ± 0.003	0.693 ± 0.016	0.701
	Gigapath (KP) [72]	0.650 ± 0.043	0.865 ± 0.032	0.563 ± 0.005	0.686 ± 0.016	0.691
	ModalTune (Ours)	$\mathbf{0.772} \pm 0.008$	0.879 ± 0.004	$\underline{0.608} \pm 0.023$	$\mathbf{0.743} \pm 0.004$	**0.750**
	ModalTune Pan-Cancer (Ours)	0.757 ± 0.039	0.860 ± 0.006	0.586 ± 0.020	0.705 ± 0.007	0.727

Table 2. Cancer subtype prediction balanced accuracy and survival prediction C-index scores across 4 cancer types. Best model in **bold**, second best is underlined. Here, LP refers to linear probing, cat refers to concatenation, and KP refers to Kronecker product.

	COADREAD	BLCA
	Cancer Subtype Prediction	
Gigapath LP	0.510	0.569
Gigapath Sup. (cat)	0.581 ± 0.006	0.703 ± 0.018
Gigapath Cls. (cat)	0.504 ± 0.030	0.497 ± 0.005
Gigapath Surv. (cat)	0.500 ± 0.000	0.497 ± 0.005
ModalTune	$\mathbf{0.574} \pm 0.024$	$\underline{0.664} \pm 0.025$
ModalTune Pan-Cancer	$\underline{0.564} \pm 0.034$	$\mathbf{0.689} \pm 0.035$
	Survival Prediction	
Gigapath LP	0.482	0.603
Gigapath Sup. (cat)	0.528 ± 0.023	0.673 ± 0.020
Gigapath Cls. (cat)	0.479 ± 0.042	0.610 ± 0.063
Gigapath Surv. (cat)	0.512 ± 0.061	0.552 ± 0.052
ModalTune	$\underline{0.539} \pm 0.068$	$\underline{0.629} \pm 0.041$
ModalTune Pan-Cancer	$\mathbf{0.543} \pm 0.020$	$\mathbf{0.672} \pm 0.046$

Table 3. Generalization study on OOD datasets using different models, compared with Gigapath Sup. (cat) trained directly on the OOD data. Best OOD model in **bold**, second best is underlined.

ever, the Pan-Cancer model is significantly more resource-efficient, requiring only one model instead of one per cancer site. As the number of sites N_{sites} and tasks T increase, scalability and efficiency benefits become more pronounced, surpassing simple baselines. This highlights a trade-off where the Pan-Cancer model enhances resource efficiency with minor performance compromises.

We hypothesize the performance drop on in-domain datasets is due to variations in task difficulty across cancer sites. Training standard ModalTune revealed variability in the optimal validation epoch per site—NSCLC reached peak performance 10 epochs earlier than GBMLGG, which required more training to converge. In the pan-cancer model, early stopping optimized for NSCLC led to underfitting for GBMLGG, while extended training overfitted NSCLC. Despite these convergence challenges reducing in-domain performance, exposure to diverse cancer sites in the pan-cancer setup greatly enhances generalizability on OOD datasets (Tab. 3). Thus, we believe the pan-cancer approach is promising and emphasize the need for optimization strategies to recover in-domain performance.

4.6. Qualitative Analysis

To assess pathway contributions and modality interactions, we analyze two breast cancer cases from high and low-risk categories, as shown in Fig. 3. See Supp. for more qualitative analysis of t-SNE and Kaplan-Meier (Sec. 14).

Pathway Contributions: We applied Integrated Gradients [60] to analyze pathway contributions as shown in Fig. 3 (e,f). In the high-risk case, incorporating pathway signals

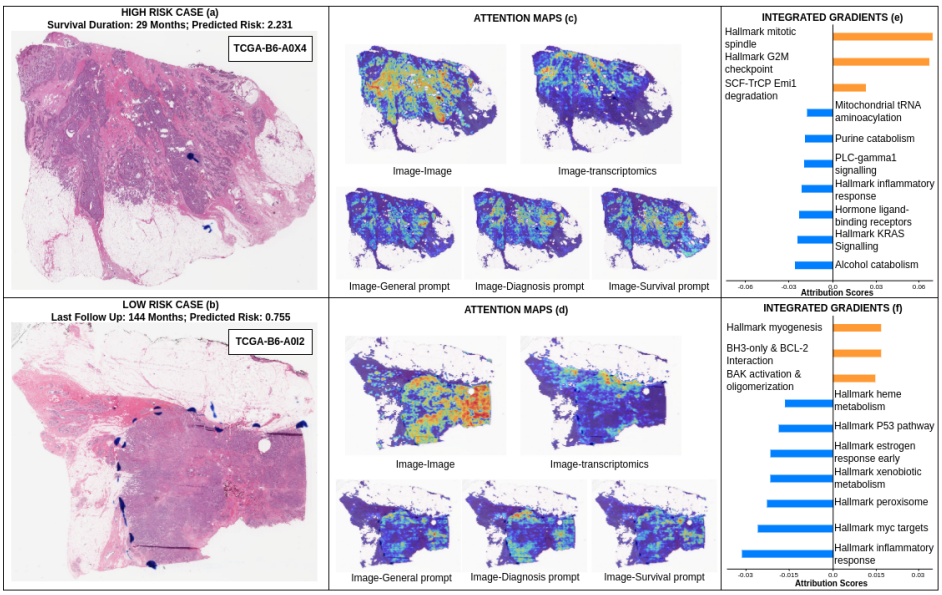

Figure 3. Qualitative analysis of breast cancer cases, highlighting a high-risk (**a**) and a low-risk case (**b**). (1) **Attention Maps (c,d)** depict cross-modal and cross-task interactions, with heatmap colors indicating importance (red: high, blue: low). (2) **Integrated Gradients (e,f)** shows the top 10 pathways influencing risk, with orange bars indicating pathways increased risk and blue bars indicating pathways decreased risk. The analysis highlights biologically relevant pathways (e.g., G2M checkpoint) and morphologically significant regions.

raised the risk score from 2.07 to 2.23. Key contributors included mitotic spindle assembly [67], G2M checkpoint regulation [42], and SCF-beta-TrCP Mediated Emi1 degradation [45], consistent with their roles in cell cycle progression and tumor proliferation. In both cases, the inflammatory response [77] lowered risk, reflecting its association with reduced oncogenic activity. Metabolic pathways like alcohol and purine catabolism [70] also reduced risk but require further study. In the low-risk case, the score dropped from 1.09 to 0.75, driven by inflammatory response [77], p53 activation [21], MYC targets [56], and estrogen signaling suppression [48]—aligning with established roles in modulating tumor progression or suppression. Interestingly, myogenesis [36] and apoptosis regulation (BH3-only & BCL-2 interaction, BAK activation, and olgomerization) contributed to increased risk, with myogenesis also linked to risk in SurvPath's analysis [31], warranting further study.

Interaction Analysis: We analyzed cross- and self- attention maps to understand different learned interactions, as shown in Fig. 3 (c,d). Specifically, we examined the self-attention of LongNet `[CLS]` tokens with patch tokens (image-image), averaged cross-attention of pathway tokens with patch tokens (image-transcriptomics), and cross-attention of patch tokens with task prompts (image-`[task]` prompt). Our findings show that image-image self-attention primarily focuses on tumor regions, capturing key pathological features, while image-transcriptomics attention highlights tumor-associated stroma. Differences in attention maps for subtype prediction and survival task prompts suggest both shared and distinct feature utilization, with general prompts integrating elements from both.

5. Conclusion

In this study, we introduced ModalTune, a flexible fine-tuning framework that maximizes dataset usage for SLFMs in histopathology. To our knowledge, ModalTune is the first multi-modal, multi-task *and* pan-cancer tuning framework. By leveraging inter-modal and inter-task interactions, ModalTune outperformed all single-task baselines while retaining generalizability. Its resource-efficient variant, ModalTune Pan-Cancer, further improves generalization on small OOD datasets, outperforming other fine-tuned models by a large margin while remaining competitive with supervised baselines. This underscores ModalTune's ability to enhance SLFMs' capacity for handling smaller datasets.

Despite its strengths, ModalTune faces two key challenges. Firstly, because the interaction between multi-modal and image features is mediated by cross-attention, it can be computationally expensive. We mitigated this by compressing pathway tokens, but future work may explore cheaper formulations of cross-attention. Secondly, the model may underfit one task while overfitting another due to varying task complexities, as seen in our pan-cancer experiments. Future work will explore strategies to mitigate the pan-cancer convergence issue, and ModalTune's adaptability to other SLFMs. Looking ahead, we also aim to incorporate additional tasks and modalities to ModalTune.

6. Acknowledegments

This work was supported by funding from the Canadian Institutes of Health Research (CIHR grant #162327), Canada Foundation for Innovation (40206), and the Ontario Research Fund. T.X. is supported by the NSERC PGS-D award and Google PhD Fellowship. A.L.M. is partially supported by the Tory Family Chair in Oncology.

References

[1] Faruk Ahmed, Andrew Sellergren, Lin Yang, Shawn Xu, Boris Babenko, Abbi Ward, Niels Olson, Arash Mohtashamian, Yossi Matias, Greg S Corrado, et al. Pathalign: A vision-language model for whole slide images in histopathology. *arXiv preprint arXiv:2406.19578*, 2024. 1, 2

[2] Ehsan Amid, Rohan Anil, Wojciech Kotłowski, and Manfred K Warmuth. Learning from randomly initialized neural network features. *arXiv preprint arXiv:2202.06438*, 2022. 5

[3] Elahe Arani, Fahad Sarfraz, and Bahram Zonooz. Noise as a resource for learning in knowledge distillation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3129–3138, 2021. 5

[4] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4

[5] Sudipto Baul, Khandakar Tanvir Ahmed, Qibing Jiang, Guangyu Wang, Qian Li, Jeongsik Yong, and Wei Zhang. Integrating spatial transcriptomics and bulk rna-seq: predicting gene expression with enhanced resolution through graph attention networks. *Briefings in Bioinformatics*, 25(4): bbae316, 2024. 3

[6] Christos Boutsidis, Anastasios Zouzias, and Petros Drineas. Random projections for k-means clustering. *Advances in neural information processing systems*, 23, 2010. 5

[7] Gabriele Campanella, Shengjia Chen, Ruchika Verma, Jennifer Zeng, Aryeh Stock, Matt Croken, Brandon Veremis, Abdulkadir Elmas, Kuan-lin Huang, Ricky Kwan, et al. A clinical benchmark of public self-supervised pathology foundation models. *arXiv preprint arXiv:2407.06508*, 2024. 1

[8] Richard J Chen, Ming Y Lu, Jingwen Wang, Drew FK Williamson, Scott J Rodig, Neal I Lindeman, and Faisal Mahmood. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Transactions on Medical Imaging*, 41(4):757–770, 2020. 2, 3, 5

[9] Richard J Chen, Ming Y Lu, Muhammad Shaban, Chengkuan Chen, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. Whole slide images are 2d point clouds: Context-aware survival prediction using patch-based graph convolutional networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 339–349. Springer, 2021. 1

[10] Richard J. Chen, Ming Y. Lu, Wei-Hung Weng, Tiffany Y. Chen, Drew F.K. Williamson, Trevor Manz, Maha Shady, and Faisal Mahmood. Multimodal Co-Attention Transformer for Survival Prediction in Gigapixel Whole Slide Images. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3995–4005. IEEE, 2021. 2, 3, 5, 7

[11] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024. 1

[12] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022. 2, 3, 4

[13] Ozan Ciga, Tony Xu, and Anne Louise Martel. Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications*, 7:100198, 2022. 1

[14] Michael Crawshaw. Multi-task learning with deep neural networks: A survey. *arXiv preprint arXiv:2009.09796*, 2020. 2, 3

[15] Cameron Davidson-Pilon. lifelines: survival analysis in python. *Journal of Open Source Software*, 4(40):1317, 2019. 6

[16] Kexin Ding, Mu Zhou, Dimitris N Metaxas, and Shaoting Zhang. Pathology-and-genomics multimodal transformer for survival outcome prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 622–631. Springer, 2023. 2

[17] Tong Ding, Sophia J Wagner, Andrew H Song, Richard J Chen, Ming Y Lu, Andrew Zhang, Anurag J Vaidya, Guillaume Jaume, Muhammad Shaban, Ahrong Kim, et al. Multimodal whole slide foundation model for pathology. *arXiv preprint arXiv:2411.19666*, 2024. 1, 2, 3, 4, 5

[18] Babak Ehteshami Bejnordi, Jimmy Lin, Ben Glass, Maeve Mullooly, Gretchen L Gierach, Mark E Sherman, Nico Karssemeijer, Jeroen van der Laak, and Andrew H Beck. Deep learning-based assessment of tumor-associated stroma for diagnosing breast cancer in histopathology images. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 929–932, 2017. 3

[19] Haitham A Elmarakeby, Justin Hwang, Rand Arafeh, Jett Crowdis, Sydney Gang, David Liu, Saud H AlDubayan, Keyan Salari, Steven Kregel, Camden Richter, et al. Biologically informed deep neural network for prostate cancer discovery. *Nature*, 598(7880):348–352, 2021. 3, 5, 7

[20] Zeyu Gao, Bangyang Hong, Yang Li, Xianli Zhang, Jialun Wu, Chunbao Wang, Xiangrong Zhang, Tieliang Gong, Yefeng Zheng, Deyu Meng, et al. A semi-supervised multi-task learning framework for cancer classification with weak annotation in whole-slide images. *Medical Image Analysis*, 83:102652, 2023. 3

[21] Milena Gasco, Shukri Shami, and Tim Crook. The p53 pathway in breast cancer. *Breast cancer research*, 4:1–7, 2002. 8

[22] Raja Giryes, Guillermo Sapiro, and Alex M Bronstein. Deep neural networks with random gaussian weights: A universal

classification strategy? *IEEE Transactions on Signal Processing*, 64(13):3444–3457, 2016. 5

[23] Mary J Goldman, Brian Craft, Mim Hastie, Kristupas Repečka, Fran McDade, Akhil Kamath, Ayan Banerjee, Yunhai Luo, Dave Rogers, Angela N Brooks, et al. Visualizing and interpreting cancer genomics data via the xena platform. *Nature biotechnology*, 38(6):675–678, 2020. 5

[24] Simon Graham, Quoc Dang Vu, Mostafa Jahanifar, Shan E Ahmed Raza, Fayyaz Minhas, David Snead, and Nasir Rajpoot. One model is all you need: multi-task learning enables simultaneous histology image segmentation and classification. *Medical Image Analysis*, 83:102685, 2023. 3

[25] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2022-June:15979–15988, 2022. 1, 2

[26] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 2

[27] Shih-Cheng Huang, Anuj Pareek, Saeed Seyyedi, Imon Banerjee, and Matthew P Lungren. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ digital medicine*, 3(1):136, 2020. 2

[28] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136, 2018. 1, 5, 7

[29] Guillaume Jaume, Lukas Oldenburg, Anurag Vaidya, Richard J Chen, Drew FK Williamson, Thomas Peeters, Andrew H Song, and Faisal Mahmood. Transcriptomics-guided slide representation learning in computational pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9632–9644, 2024. 1

[30] Guillaume Jaume, Lukas Oldenburg, Anurag Vaidya, Richard J. Chen, Drew F. K. Williamson, Thomas Peeters, Andrew H. Song, and Faisal Mahmood. Transcriptomics-guided slide representation learning in computational pathology-CVPR 2024. https://github.com/mahmoodlab/TANGLE, 2024. [Online; accessed 2025-03-07]. 6, 4

[31] Guillaume Jaume, Anurag Vaidya, Richard J Chen, Drew FK Williamson, Paul Pu Liang, and Faisal Mahmood. Modeling dense multimodal interactions between biological pathways and histology for survival prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11579–11590, 2024. 2, 3, 5, 7, 8

[32] Guillaume Jaume, Anurag Vaidya, Andrew Zhang, Andrew H. Song, Richard J. Chen, Sharifa Sahai, Dandan Mo, Emilio Madrigal, Long Phi Le, and Faisal Mahmood. Multistain pretraining for slide representation learning in pathology. In *European Conference on Computer Vision*, pages 19–37. Springer, 2024. 1

[33] Ting Jin, Xingran Xie, Renjie Wan, Qingli Li, and Yan Wang. Gene-induced multimodal pre-training for image-omic classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 508–517. Springer, 2023. 2

[34] William B Johnson, Joram Lindenstrauss, et al. Extensions of lipschitz mappings into a hilbert space. *Contemporary mathematics*, 26(189-206):1, 1984. 5

[35] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. *Advances in neural information processing systems*, 30, 2017. 5, 7

[36] Robert S Krauss, Francesca Cole, Ursula Gaio, Giichi Takaesu, Wei Zhang, and Jong-Sun Kang. Close encounters: regulation of vertebrate skeletal myogenesis by cell-cell contact. *Journal of cell science*, 118(11):2355–2362, 2005. 8

[37] Ritika Kundra, Hongxin Zhang, Robert Sheridan, Sahussapont Joseph Sirintrapun, Avery Wang, Angelica Ochoa, Manda Wilson, Benjamin Gross, Yichao Sun, Ramyasree Madupuri, et al. Oncotree: a cancer classification system for precision oncology. *JCO clinical cancer informatics*, 5: 221–230, 2021. 5, 1

[38] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2

[39] Lingqiao Li, Xipeng Pan, Huihua Yang, Zhenbing Liu, Yubei He, Zhongming Li, Yongxian Fan, Zhiwei Cao, and Longhao Zhang. Multi-task deep learning for fine-grained classification and grading in breast cancer histopathological images. *Multimedia Tools and Applications*, 79:14509–14528, 2020. 3

[40] Zheng Li, Xiang Li, Xinyi Fu, Xin Zhang, Weiqiang Wang, Shuo Chen, and Jian Yang. Promptkd: Unsupervised prompt distillation for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26617–26626, 2024. 5

[41] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2

[42] Markus Löbrich and Penny A Jeggo. The impact of a negligent g2/m checkpoint on genomic instability and cancer induction. *Nature Reviews Cancer*, 7(11):861–869, 2007. 8

[43] 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. 1, 2, 5

[44] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Melissa Zhao, Aaron K Chow, Kenji Ikemura, Ahrong Kim, Dimitra Pouli, Ankush Patel, et al. A multimodal generative ai copilot for human pathology. *Nature*, 634(8033): 466–473, 2024. 2

[45] Florence Margottin-Goguet, Jerry Y Hsu, Alexander Loktev, Harn-Mei Hsieh, Julie DR Reimann, and Peter K Jackson. Prophase destruction of emi1 by the scfβtrcp/slimb ubiquitin ligase activates the anaphase promoting complex to allow progression beyond prometaphase. *Developmental cell*, 4(6): 813–826, 2003. 8

[46] Pooya Mobadersany, Safoora Yousefi, Mohamed Amgad, David A Gutman, Jill S Barnholtz-Sloan, José E Velázquez Vega, Daniel J Brat, and Lee AD Cooper. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences*, 115(13):E2970–E2979, 2018. 3, 5

[47] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 5

[48] Masanori Oshi, Yoshihisa Tokumaru, Fernando A Angarita, Li Yan, Ryusei Matsuyama, Itaru Endo, and Kazuaki Takabe. Degree of early estrogen response predict survival after endocrine therapy in primary and metastatic er-positive breast cancer. *Cancers*, 12(12):3557, 2020. 8

[49] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 6

[50] Linmin Pei, Wei-Wen Hsu, Ling-An Chiang, Jing-Ming Guo, Khan M Iftekharuddin, and Rivka Colen. A hybrid convolutional neural network based-method for brain tumor classification using mmri and wsi. In *6th International Workshop, BrainLes 2020, International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 487–496. Springer, 2021. 2

[51] Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. Adapterhub: A framework for adapting transformers. *arXiv preprint arXiv:2007.07779*, 2020. 2

[52] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. *Proceedings of Machine Learning Research*, 139:8748–8763, 2021. 2

[53] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 5

[54] Vishwesh Ramanathan, Pushpak Pati, Matthew McNeil, and Anne L Martel. Ensemble of prior-guided expert graph models for survival prediction in digital pathology. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 262–272. Springer, 2024. 1

[55] Andrew M Saxe, Pang Wei Koh, Zhenghao Chen, Maneesh Bhand, Bipin Suresh, and Andrew Y Ng. On random weights and unsupervised feature learning. In *ICML*, page 6, 2011. 5

[56] Amy Schulze, Masanori Oshi, Itaru Endo, and Kazuaki Takabe. Myc targets scores are associated with cancer aggressiveness and poor survival in er-positive primary and metastatic breast cancer. *International journal of molecular sciences*, 21(21):8127, 2020. 8

[57] George Shaikovski, Adam Casson, Kristen Severson, Eric Zimmermann, Yi Kan Wang, Jeremy D Kunz, Juan A Retamero, Gerard Oakley, David Klimstra, Christopher Kanan, et al. Prism: A multi-modal generative foundation model for slide-level histopathology. *arXiv preprint arXiv:2405.10254*, 2024. 2

[58] Wei Shao, Hang Shi, Jianxin Liu, Yingli Zuo, Liang Sun, Tiansong Xia, Wanyuan Chen, Peng Wan, Jianpeng Sheng, Qi Zhu, and Daoqiang Zhang. Multi-Instance Multi-Task Learning for Joint Clinical Outcome and Genomic Profile Predictions From the Histopathological Images. *IEEE Transactions on Medical Imaging*, 43(6):2266–2278, 2024. 3

[59] 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

[60] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017. 7

[61] Kevin Thandiackal, Boqi Chen, Pushpak Pati, Guillaume Jaume, Drew FK Williamson, Maria Gabrani, and Orcun Goksel. Differentiable zooming for multiple instance learning on whole-slide images. In *European Conference on Computer Vision*, pages 699–715. Springer, 2022. 1

[62] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. MLP-Mixer: An all-MLP Architecture for Vision. *Advances in Neural Information Processing Systems*, 29:24261–24272, 2021. 3, 5

[63] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. Review the cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary Oncology/Współczesna Onkologia*, 2015(1):68–77, 2015. 2, 5

[64] Anurag Vaidya, Andrew Zhang, Guillaume Jaume, Andrew H Song, Tong Ding, Sophia J Wagner, Ming Y Lu, Paul Doucet, Harry Robertson, Cristina Almagro-Perez, et al. Molecular-driven foundation model for oncologic pathology. *arXiv preprint arXiv:2501.16652*, 2025. 1

[65] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[66] Roel GW Verhaak, Katherine A Hoadley, Elizabeth Purdom, Victoria Wang, Yuan Qi, Matthew D Wilkerson, C Ryan Miller, Li Ding, Todd Golub, Jill P Mesirov, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1. *Cancer cell*, 17(1):98–110, 2010. 6

[67] Claire E Walczak and Rebecca Heald. Mechanisms of mitotic spindle assembly and function. *International review of cytology*, 265:111–158, 2008. 8

[68] Wenhui Wang, Shuming Ma, Hanwen Xu, Naoto Usuyama, Jiayu Ding, Hoifung Poon, and Furu Wei. When an image is

worth 1,024 x 1,024 words: A case study in computational pathology. *arXiv preprint arXiv:2312.03558*, 2023. 3

[69] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical image analysis*, 81:102559, 2022. 1

[70] RWE Watts. Molecular variation in relation to purine metabolism. *Journal of Clinical Pathology. Supplement (Royal College of Pathologists).*, 8:48, 1974. 8

[71] Wei-Hung Weng, Yuannan Cai, Angela Lin, Fraser Tan, and Po-Hsuan Cameron Chen. Multimodal multitask representation learning for pathology biobank metadata prediction. *arXiv preprint arXiv:1909.07846*, 2019. 3

[72] 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. 1, 2, 3, 5, 7

[73] Yingxue Xu and Hao Chen. Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21241–21251, 2023. 3

[74] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 2

[75] Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, et al. Ultramedical: Building specialized generalists in biomedicine. *Advances in Neural Information Processing Systems*, 37:26045–26081, 2025. 4

[76] Bo Zhao and Hakan Bilen. Dataset Condensation with Distribution Matching. *Proceedings - 2023 IEEE Winter Conference on Applications of Computer Vision, WACV 2023*, pages 6503–6512, 2023. 5

[77] Ruijun Zhao, Chaoyu Xie, Yu Gong, Songzhi Wei, Mei Yuan, Jinfeng Gan, and Wenyan Chen. A novel inflammatory response-related gene signature predicts immune status and prognosis of breast cancer. *Journal of Oncology*, 2022 (1):5468858, 2022. 8

[78] Weiqin Zhao, Shujun Wang, Maximus Yeung, Tianye Niu, and Lequan Yu. Mulgt: Multi-task graph-transformer with task-aware knowledge injection and domain knowledge-driven pooling for whole slide image analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3606–3614, 2023. 3

[79] Yi Zheng, Rushin H Gindra, Emily J Green, Eric J Burks, Margrit Betke, Jennifer E Beane, and Vijaya B Kolachalama. A graph-transformer for whole slide image classification. *IEEE transactions on medical imaging*, 41(11):3003–3015, 2022. 1

[80] Qifeng Zhou, Wenliang Zhong, Yuzhi Guo, Michael Xiao, Hehuan Ma, and Junzhou Huang. Pathm3: A multimodal multi-task multiple instance learning framework for whole slide image classification and captioning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 373–383. Springer, 2024. 2, 3